





# Spatio–Temporal Attention-Based Deep Learning Framework for Mesoscale Eddy Trajectory Prediction

Xuegong Wang , Chong Li , Xinning Wang , Lining Tan, and Jin Wu 

**Abstract**—Accurate prediction of mesoscale eddy trajectories requires efficient models with large-size available data instances to capture the main eddy characteristics. However, there is a lack of salient attention mechanisms that can recognize the demand of extracting the aggregated features over multidimensional eddy data sources. Additionally, deep learning techniques are very important for eddy trajectory prediction to dynamically capture properties of mesoscale eddies in the South China Sea. In this article, we propose a spatio–temporal attention-based deep learning framework that can orchestrate heterogeneous data integration and propagation trajectory forecast together. It consists of a novel autoencoder equipped with channel and spatial attention mechanisms (CSA-encoder), and a gated recurrent unit (GRU) network with temporal attention layer (TA-GRU). CSA-encoder compresses stereoscopic eddy data with convolutional layers and generates the small-scale and high-quality dataset as the input of TA-GRU. The finer grained TA-GRU method is extended to accurately predict eddy trajectories with more valuable imagery information so that the temporal attention mechanism can automatically select relevant regions within the next 14 days. Our cross-validation results demonstrate that our framework averagely achieves a lower distance error (9 km) and 54% performance improvement over the baseline GRU technique in the next one day, and outperforms two state-of-the-art techniques of long short-term memory and recurrent neural network by 54.9% and 65.6%, respectively.

**Index Terms**—Attention mechanism, autoencoder, gated recurrent unit network, mesoscale eddy, trajectory prediction.

## I. INTRODUCTION

THE mesoscale eddy has proven itself as a viable ocean phenomenon for a better understanding of heat and mass transfer, nonlinear energy transport, global climate change, and marine resource distribution [1]–[4]. It also has a significant influence on annual subduction rate within the main thermocline [5] and phytoplankton dynamics in the low sea level [6]. Hence, there is an emerging need of precisely predicting mesoscale eddies' trajectories in advance, so that the

Manuscript received August 31, 2021; revised December 7, 2021, January 17, 2022, and March 9, 2022; accepted May 7, 2022. Date of publication May 11, 2022; date of current version May 23, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171420, in part by the Natural Science Foundation of Shandong under Grant ZR201910230031, and in part by Fundamental Research Funds for Central Universities under Grant 202213016. (Corresponding author: Xinning Wang.)

Xuegong Wang, Chong Li, and Xinning Wang are with the Ocean University of China, Qingdao 266100, China (e-mail: xuegong\_wang@stu.ouc.edu.cn; lichong7332@ouc.edu.cn; wangxinning@ouc.edu.cn).

Lining Tan is with the Xi'an Research Institute of High Technology, Xi'an 710000, China (e-mail: tg.dracula@gmail.com).

Jin Wu is with the Hong Kong University of Science and Technology, Hong Kong (e-mail: jin\_wu\_uestc@hotmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3174239

safe navigation of offshore platforms, submarines, and fishing boats can precede the appearance of environmental hazards induced by powerful eddies [7], [8]. Furthermore, forecasting eddy trajectories can help marine scientists analyze the seasonal and interannual variability of characteristics beforehand, so *in situ* observation instruments are deployed to track and monitor a particular eddy from its generation to disappearing.

A major challenge for eddy trajectory prediction is posed by data manipulation of heterogeneous eddy data [9]–[11], including multidimensional data processing, high-density data compression, and heterogeneous data combination. Large-scale eddy data residing in different sources make traditional data mining mechanisms unable to fully comprehend relationships in mesoscale eddy data. Furthermore, due to the satellite-based observation limit, dynamic model prediction capacity, such as numerical simulation and statistical analysis, may diminish as the marked increase of high-resolution data [12]–[15]. To achieve high forecasting performance, such conventional methodologies are mainly based on representative physical models and massive heterogeneous eddy data to make features comprehensive, which increase the consumption of computation resources. In addition, they require different types of eddy data to analyze sufficient mesoscale eddy attributes [16]. But multidimensional eddy data, including sea level anomaly (SLA), water temperature (WT), sea surface height (SSH), and sea surface temperature (SST), which cover sea surface and undersea parts of the whole eddy, have not been integrated or refined to enhance trajectory prediction performance.

Compared with aforementioned techniques, deep learning-based approaches allow computational layers that are composed of multiple processing neurons to automatically study representations of eddy data. Hence deep learning-based models are proposed to learn the hidden characteristics for tracking [17]–[19] or identifying mesoscale eddies over a short period of time, such as PSPNet [20], Deepeddy [21], and OEDNet [22]. Such models only supervisedly learn eddy features with a single dataset on the surface of the ocean (either SLA data or SAR images) instead of integrating heterogeneous eddy data. Therefore, how to design a reasonable data management strategy that integrates heterogeneous eddy data and unifies high-dimensional eddy features becomes the key issue to be addressed.

Another challenge for time-sequence trajectory prediction is that current techniques cannot capture the correlation between multiple eddy characteristics over a long period of time [23], [24], lacking an efficient stage to focus on the important vertical and horizontal areas of mesoscale eddy data blocks. Since the

trajectories of eddy are time sequences, neural networks which have a capability of memorizing data of previous moments are essential for the time-sequence prediction, such as RNN [25], LSTM [26], and GRU [27]. The LSTM algorithm is utilized to predict the trajectories and properties of eddies with only few eddy physical characteristics from single 1-D dataset, which has a lack of other higher dimensional datasets for supplementing adequate mesoscale eddy information [28]. The Conv-LSTM is also applied to nowcast the evolution of eddies with SLA data [29], but without the spatial attention mechanism so that this network cannot focus on key eddy areas or ignore irrelevant non-eddy attributes importing an adverse effect on network convergence. This challenge demands an attention mechanism that can be in favor of deep learning-based models for high accuracy of eddy trajectory prediction. Though there are many works on attention-based deep learning strategies for classification and recognition [30]–[33], they are all oblivious of refinement of stereoscopic eddy data.

To be specific, for these existing both traditional and deep learning methods, eddy prediction faces several challenges in the following aspects.

- 1) *Stereoscopic Structure*: Different from other sea surface phenomena, mesoscale eddy is stereoscopic with a diameter of hundreds of kilometers and a depth of more than thousands of kilometers. But existing eddy prediction models just exploit a single sea-surface dataset without considering its stereoscopic features, which makes it difficult to recognize mesoscale eddies by inadequate features.
- 2) *Heterogeneous Datasets*: Although there are heterogeneous eddy datasets (e.g., SLA and WT) to characterize and evaluate the process of eddy activities, no methods can satisfy the temporospatial requirement of multiple eddy data combination. There need to be more datasets imported and fused to describe eddy's stereoscopic structure. But the critical eddy features are submerged with massive irrelevant information in data combination, which may introduce distraction on information processing and eddy prediction.
- 3) *Continuous Time Sequences*: Mesoscale eddy data include both spatial and temporal information in continuous time sequences. For different eddy prediction networks, the data at different time steps can produce different efforts on eddy forecasting performance due to other insignificant temporospatial characteristics. And the temporal characteristics of mesoscale eddies can be weakened by unimportant step data and the difficulty of predicting eddies can be aggravated.

Theoretically, the network's ability to focus on key eddy areas and time steps can be realized by spatial and temporal attention mechanism. Recently, to detect temporal and spatial characteristics, the attention mechanism as a new technique has been applied to predict SST, SSH, and ocean current. Equipped with time-step attention mechanism, LSTM method can obtain better prediction results with temporal correlations of SST data than traditional LSTM [34]. In [35], self-attention mechanisms are used to discover features of ocean current time steps for improving forecast accuracy rate in the next ten days. The majority

of time-based attention mechanisms contribute to helping neural network pay more attention to more important time characteristics, ignoring other significant properties, such as diverse spatial features. An LSTM model for SSH prediction [36] attempts to import both time and space attention mechanisms, but just horizontal ocean surface data samples are used instead of underwater data instances. We observe that both time and space attention mechanisms may carry high prediction accuracy, but for eddy time data, corresponding spatio-temporal attention mechanisms have not been exploited. Hence we devise spatio-temporal attention module to automatically learn the key contributions of heterogeneous eddy data covering eddy's stereoscopic structure, including original trajectory data, sea level anomalies, and water temperatures for precisely forecasting different eddy types.

For overcoming these aforementioned problems, it is imperative to integrate attention mechanisms with deep learning approaches with a comprehensive dataset. To this end, we propose a spatio-temporal attention-based deep learning framework to improve the forecasting performance of eddy trajectories with heterogeneous eddy data. Based on the observation that the time and space attentions improve the forecasting quality, our framework exhibits a complementary effect that contributes to multidimensional eddy data integration with channel and spatial attention modules. It also traverses and downloads diverse datasets from three ocean institutions indicating detailed coverage of mesoscale eddies from sea surface to sea bottom. In addition, temporal attention-based GRU network dynamically concentrates on the key eddy attributes.

In summary, the main contributions of this article are summarized as follows.

- 1) In order to solve the first challenge caused by stereoscopic eddy structure, we combine heterogeneous eddy features including sea surface and undersea data, and construct a new multidimensional dataset to cover the whole stereoscopic structure of eddy, for subsequently predicting eddy trajectories in the SCS. A data processing module automatically downloads and combines heterogeneous eddy data from three main institutions, including eddy trajectory data, SLA data, and WT data. In addition, it efficiently extracts, concatenates, and processes the eddy center area data from multiple satellite missions in the SCS.
- 2) We propose an attention-based eddy trajectory forecasting framework to overcome the second and third challenges, mainly including a channel and spatial attention-based autoencoder (CSA-encoder) and a temporal attention-based GRU network (TA-GRU), to further focus on critical eddy regions and time steps, incorporating and aggregating spatio-temporal eddy features. By taking attention mechanisms into trajectory prediction, more key information can be refined from multiple eddy data and the effectiveness and performance of trajectory prediction are also improved.
- 3) We evaluate the spatio-temporal attention-based mechanism and analyze the prediction performance in comparison to other deep learning-based methods. Experimental results illustrate our novel framework achieves the lowest daily center error of 9 km on average, and preserves a

lower center error for next 14-day forecasting. Compared to traditional methods, such as RNN, LSTM, and GRU, our method obtains approximately 65%, 56%, and 53% accuracy improvements.

The rest of this article is organized as follows. Section II summarizes the related work, including prediction networks without attention mechanisms and prediction networks with attention mechanisms. In Section III, data construction, structure of CSA-encoder and TA-GRU are explained in detail. Section IV evaluates the data compression results of CSA-encoder and prediction performance of TA-GRU network, respectively. Finally, Section V concludes this article.

## II. RELATED WORK

### A. Prediction Networks Without Attention Mechanisms

The deep learning networks, such as RNN, LSTM, and GRU, can memorize the previous data and their moments for time sequence problems. However, there are some differences among the prediction networks. For example, RNN method can calculate the current cell state with the output of previous cells, but its chain derivation causes gradient exploding and vanishing when involving long sequence input [25]; LSTM alleviates RNN's gradient problems, using three gates (input, forget, and output gates) to balance the current input and previous state in the cell [26]; GRU simplifies LSTM structure with two gates (update and reset gates), bringing nearly a 1/3 parameter reduction [27].

To improve the prediction performance of sea surface data, several studies designed time-series predicting networks, based on RNN, LSTM, and GRU's memorizing capability. For instance, Song *et al.* merged ResNet and LSTM for enhancing SLA prediction performance [37]. Chen *et al.* [39] combined LSTM and migration learning method for ocean image processing [38], and Fan *et al.* utilized CNN+LSTM to apply acceleration of turbulence flow simulation. Ma *et al.* [29] proposed a Conv-LSTM network to predict SSH data and then obtained approximately 60% matching eddies on the next seventh day. Guo *et al.* [40] applied ConvGRU to design encoder-decoder model for nowcasting convective weather based on radar data. Xie *et al.* [41] also used dynamic GRU-based encoder-decoder to achieve predicting future SST code, aiming to solve the long-scale dependence problem. Based on LSTM and extra trees (ET) algorithms, Wang *et al.* [28] proposed a model for mesoscale eddy property prediction with 1-D eddy property data. Song *et al.* [42] proposed a dual path GRU method for sea surface salinity prediction to improve detection accuracy in the 14 days. Liu *et al.* [43] incorporated the social force model into LSTM network (SFM-LSTM), whose loss function was reconstructed by offset distance and direction, leading to more robust vessel trajectory prediction in different water areas. In addition, LSTM and GRU methods are also applied in the fields of the trajectory prediction of hurricanes [44], sea ice [45], and oceanic flows [46], since they can capture and memorize the time information hidden in data sequences.

But all the aforementioned studies just adopted 1-D or 2-D data on the ocean surface [47]–[49], ignoring undersea data, which contain more features than surface data. Especially, due to the mesoscale eddy's stereoscopic structure, its underwater

scale is ten times greater than the sea surface scale [50], [51]. A data fusion approach is required to obtain the comprehensive eddy data with both high spatial resolution and high temporal frequency for mesoscale eddy detection.

### B. Prediction Networks With Attention Mechanisms

To capture critical spatial or temporal information, respectively, space and time attention mechanisms are proposed for improving the performance and effectiveness [52], [53]. For space attention mechanism, there are channel and spatial attention layers. Channel attention layer applies maxpooling layer, avgpooling layer, and multilayer perceptron (MLP) to generate 1-D channel attention vector for three channels of color images, and similarly, spatial attention layer concatenates maxpooling and avgpooling layers, and utilizes MLP to obtain 2-D spatial feature map for all image pixels [54]. Temporal attention layer introduces temporal weight vectors into time-series predicting networks, assigning different attention weight values to different hidden states [55]. And the combination of channel, spatial and temporal attention mechanisms can capture the hidden spatio-temporal features comprehensively in object predicting and tracking [56]. However, current oceanographic prediction techniques rarely combine space and time characteristics with attention mechanisms [57]. Though mesoscale eddies have prolific spatial and temporal attributes, there is a salient lack of the application of spatial and temporal attention mechanisms in the fields of eddy recognition and prediction [58], [59].

Recently, several spatio-temporal attention mechanisms have since been developed to classify or predict sea surface data, such as SSH and SST. Feng *et al.* [60] applied time attention mechanism and temporal convolutional network to construct full-feature and partial-feature prediction models for large-scale SST data, achieving similar accuracy with less data. Based on spatial attention mechanism, Ren *et al.* [30] proposed a dual-attention U-Net for pixel-level segmentation of sea ice and open water, with better classification results than the original U-Net. Thongniran *et al.* [61] combined space-attentional CNN and GRU, to accurately forecast ocean currents with HF radar dataset. Liu *et al.* [36] added time and space attention mechanisms into LSTM and assigned reasonable weights for the data at each time step, which improved the accuracy of SSH prediction and proved the feasibility of attention modules. In addition, there are also attention-based methods for tropical cyclone track prediction [62] and ocean front detection [63], indicating the significant accuracy improvements by either temporal or spatial attention mechanism.

However, the majority of attention-based techniques just exploited a spatial or temporal attention mechanism, while spatio-temporal attention mechanism can lead to more accuracy improvement theoretically and empirically. For mesoscale eddy prediction, attention mechanism has not been used to enhance the detection accuracy and effectiveness. Hence to further obtain higher eddy forecasting accuracy, our predicting framework first integrates channel, spatial and temporal attention mechanisms, and focuses on critical eddy's vertical and horizontal regions according to the principal time steps. Furthermore, our framework further optimizes channel and spatial attention mechanisms to



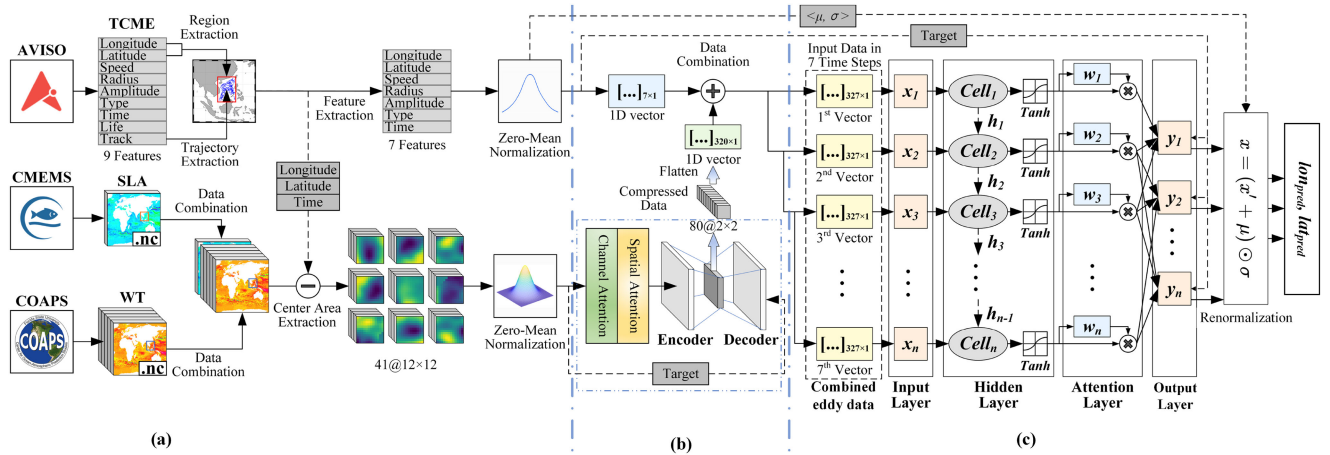


Fig. 1. Flow graph of spatio-temporal attention-based deep learning framework for eddy trajectory prediction. (a) Multisource data construction part extracts, combines, and processes heterogeneous datasets in the SCS from three institutions. (b) CSA-encoder utilizes attention mechanisms, encoder, and decoder to identify, compress, and restore eddy characteristics. (c) TA-GRU executes iterative prediction for eddy trajectories, with temporal attention layer implemented. (a) Multi-source data construction. (b) Data compression of CSA-encoder. (c) TA-GRU trajectories prediction network.

make them more suitable for eddy datasets. In summary, most prediction models just exploit the single sea surface data rather than the combination of heterogeneous data, and these techniques ignore the integration of spatial and temporal attention mechanisms that can acquire some critical spatio-temporal characteristics.

### III. METHODS

In this section, we design and implement the deep learning-based framework shown in Fig. 1 for mesoscale trajectory prediction. In Fig. 1(a), for collected TCME, SLA, and WT datasets from different institutions, the data extraction part in Section III-A withdraws the SCS's regional information from TCME and the combined data of SLA and WT. Then the refined data instances are processed by zero-mean normalization. We can observe that in Fig. 1(b) the CSA-encoder designed in Section III-B achieves data compression of SLA and WT, whose channel and spatial attention mechanisms can make the learning module concentrate on more distinctive areas. Additionally, TA-GRU with temporal attention layer in Fig. 1(c) proposed in Section III-C iteratively trains a trajectory prediction model with processed trajectory sequence samples, and finally applies renormalization to the intermediate results.

#### A. Data Construction

1) *Heterogeneous Eddy Data*: In our study, heterogeneous multidimensional datasets are involved. The datasets include SLA, WT, and trajectory characteristics of mesoscale eddies (TCME), which are shown in Fig. 1(a). SLA is provided by Copernicus Marine Environment Monitoring Service (CMEMS).<sup>1</sup> WT is from the Center for Ocean-Atmospheric Prediction Studies (COAPS),<sup>2</sup> and TCME is downloaded from

TABLE I  
FEATURES DESCRIPTION OF TCME

Feature	Description	Unit
Longitude	Longitude observation of eddies' center	°
Latitude	Latitude observation of eddies' center	°
Speed	Average speed of eddies' contour	cm/s
Radius	Radius of a circle contour area	km
Amplitude	Height difference between eddies' center and contour	cm
Type	+1 means cyclonic, -1 means anticyclonic	-
Time	Time delta from January 1, 1950	day
Life	Life time from eddy start	day
Track	Eddy identification number	-

archiving, validation, and interpretation of satellite oceanographic data (AVISO).<sup>3</sup> All of three datasets are stored in NetCDF files.

There are 624 465 global eddy trajectory records contained by TCME, consisting of nine features, such as longitude, latitude, speed, radius, amplitude, type, time, life, and track. Table I describes these features in detail. All datasets have the same temporal resolution of one day, but own different time ranges. SLA's time range is from January 1st 1993 to December 27th 2018, WT is from January 1994 to December 2015, and TCME is from January 1993 to September 2019. In consideration of different time spans of these datasets, the eddy data in the overlapping time period are chosen for training and testing, which is from January 1994 to December 2015. The sizes of TCME, SLA, and WT in the SCS are approximately 19.1 MB, 3.81 GB, and 31.9 GB, respectively.

The spatial resolution of SLA is  $1/4^\circ$ , and WT's is  $1/12^\circ$ . WT has 40-layer vertical data with the depth from 0 to 5000 m, while SLA has only one-layer data on the surface of the sea. In [64], we can see that eddies own a stereoscopic structure, indicating that eddies not only exist on the surface of the ocean but they can also extend to a depth over 2000 m. That is the reason why the WT data are chosen to supplement SLA and TCME datasets

<sup>1</sup>[Online]. Available: <https://marine.copernicus.eu/>

<sup>2</sup>[Online]. Available: <https://www.hycom.org/>

<sup>3</sup>[Online]. Available: <https://www.aviso.altimetry.fr/en/home.html>

TABLE II  
CORRESPONDENCE BETWEEN LAYER AND DEPTH FOR WT DATA IN MULTIPLE LAYERS

Layer	Depth (m)	Layer	Depth (m)	Layer	Depth (m)	Layer	Depth (m)
1	0	11	30	21	125	31	800
2	2	12	35	22	150	32	900
3	4	13	40	23	200	33	1,000
4	6	14	45	24	250	34	1,250
5	8	15	50	25	300	35	1,500
6	10	16	60	26	350	36	2,000
7	12	17	70	27	400	37	2,500
8	15	18	80	28	500	38	3,000
9	20	19	90	29	600	39	4,000
10	25	20	100	30	700	40	5,000

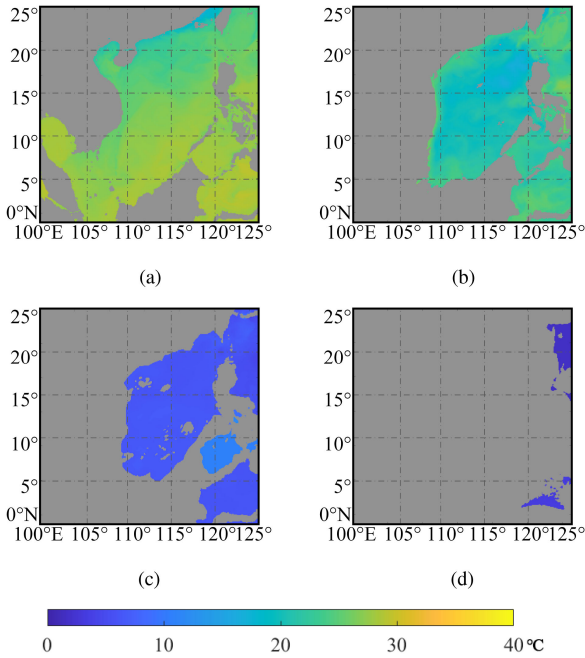


Fig. 2. WT data variation in the SCS on January 1st 1993. The color from blue to yellow indicates a temperature range of 0–40 °C, and the gray area means uncollected WT data points. (a) 10-layer WT data. (b) 20-layer WT data. (c) 30-layer WT data. (d) 40-layer WT data.

in our experiment. Table II shows the relationship of the number and depth of these 40 layers. We can find that data density is not uniform, as the data density becomes smaller and the distance interval becomes larger. WT from 1 to 20 layers cover the data of the depth [0, 100] with the interval of 2 or 5 m. In our study, layers 21–33 cover the water temperature of 125–1000 m, and 34–40 layers include the water imagery of 1000–5000 m. The variations of 10, 20, 30, and 40-layer WT data are shown in Fig. 2, corresponding to 25, 100, 700, and 5,000 m depth in Table II, respectively, where the gray area means uncollected data points. From Fig. 2(a)–(d), water temperature in the SCS drops from 25.94 °C to 2.78 °C, and uncollected area, caused by sampling costs and submarine topography, gradually expands with the water depth increasing. Evidently, the water temperature of the SCS nearly decreases to 0 °C with 40-layer water data and data collection has pertinent issues and challenges so that the uncollected area is enlarged in Fig. 2(c) and (d).

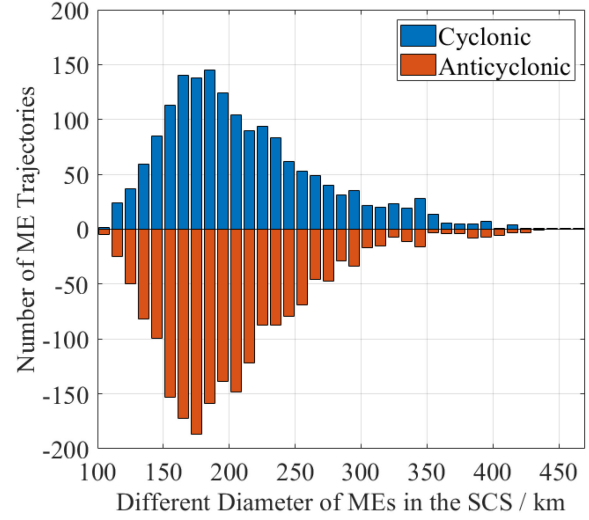


Fig. 3. Diameter distribution of cyclonic and anticyclonic eddy trajectories in the SCS from January 1993 to September 2019. Both of them conform the Rayleigh distribution, whose peak value are between 160 and 180 km. The number of anticyclonic eddies is 1928, higher than the cyclonic (1669).

TABLE III  
EDDY DIAMETER DISTRIBUTION TABLE (JANUARY 1993 TO SEPTEMBER 2019 IN THE SCS)

Diameter (km)	Cyclonic	Anticyclonic	Total
100–200	971 (26.99%)	1219 (33.89%)	2190 (60.88%)
200–300	559 (15.54%)	617 (17.15%)	1176 (32.69%)
300–400	128 (3.56%)	81 (2.25%)	209 (5.81%)
>400	11 (0.31%)	11 (0.31%)	22 (0.62%)
Total	1669 (46.40%)	1928 (53.60%)	3597 (100%)

2) *Eddy Data Processing*: In order to avoid overfitting in neural networks for trajectory prediction in the SCS, we expand the geographic scope with the range of 0°–25°N and 100°–125°E. In TCME dataset of Fig. 1(a) provided by AVISO, all original global mesoscale eddy records (including nine features detailed in Table I) are mixed disorderly. So first, we iterate over all records and judge whether eddy center is within above geographic region. Each trajectory in the SCS is identified according to track feature, with seven features selected to be input: longitude, latitude, speed, radius, amplitude, type, and time. Meanwhile, we choose the center location (longitude, latitude) in the next day as the ground truth for our framework. In particular, the time feature of mesoscale eddies starts from January 1st, 1950, and its format is converted to the number of days for highlighting seasonal features of eddies.

Because of different spatial resolutions of SLA (1/4°) and WT (1/12°), extra WT data points are filtered out to set its resolution to 1/4°, making their horizontal data matched. Although they have different vertical layers (1 and 40), these horizontal-matched data can be combined directly to generate 41-layer data. In addition, 12×12 SLA and WT grid data points around eddy centers are collected along with the longitude, latitude and time of eddy centers provided by TCME, covering a geographic area of 340 × 340 km approximately. In our diameter distribution analysis, given by Fig. 3 and Table III, the percentage

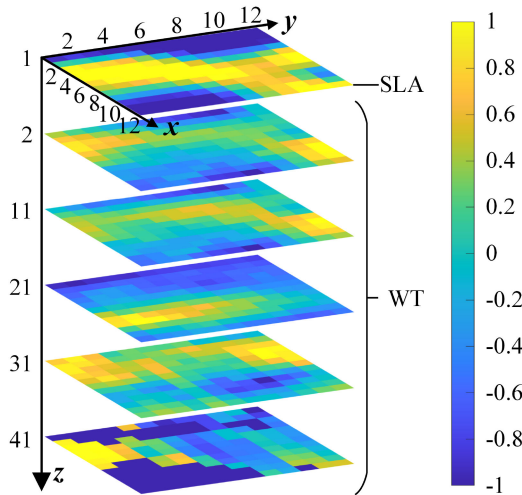


Fig. 4. Visualization of the new data covering the whole eddy with the size of  $41 \times 12 \times 12$ , including 1-layer SLA and 40-layer WT normalized data.

of eddy diameters of less than 300 km approximately reaches 93.57%. Therefore, we select  $12 \times 12$  grid points as the main representation of eddy properties. The data are purified for SLA and WT, and we convert -2 147 483 647 of SLA and -30 000 of WT into -1, to avoid a performance loss of deep learning networks, such as accuracy reduction and low convergence rate.

To integrate the three different datasets of various eddy features, normalization is required for removing the variations that affect data-processing efficiency. Here 1-D-zero-mean normalization is used to tackle anomaly values in TCME, and to build a high-quality data collection, 3-D-zero-mean normalization method for 3-D data is designed in SLA and WT data collections as follows:

$$\mu = \frac{1}{N_s N_i N_j N_k} \sum_{s=1}^{N_s} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \sum_{k=1}^{N_k} x_{s,i,j,k} \quad (1)$$

$$\sigma = \sqrt{\frac{\sum_{s=1}^{N_s} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \sum_{k=1}^{N_k} (x_{s,i,j,k} - \mu)^2}{N_s N_i N_j N_k}} \quad (2)$$

$$x'_{s,i,j,k} = \frac{x_{s,i,j,k} - \mu}{\sigma} \quad (3)$$

where  $\mu$  is the mean value of all original 3-D data and  $\sigma$  means standard deviation.  $s$  represents sample index,  $i$ ,  $j$ , and  $k$  are indexes of three dimensions (channel, height, and width).  $N_s$ ,  $N_i$ ,  $N_j$ , and  $N_k$  correspond to the total number of above four indexes separately.  $x_{s,i,j,k}$  is the data point located at channel  $i$ , height  $j$ , width  $k$  in the sample  $s$ , and  $x'_{s,i,j,k}$  stands for a normalized data point.  $x_{s,i,j,k}$  is subtracted by  $\mu$  to make new mean value equal to 0, and then divided by  $\sigma$  to ensure that new standard deviation is 1. Newly generated data satisfy the normal distribution condition after two zero-mean normalizations.

After abovementioned data processing and construction of SLA and WT data, the visualization of our layered data for one eddy is shown in Fig. 4, whose size is  $41 \times 12 \times 12$ . Layer 1 stands for  $12 \times 12$  SLA data and the remaining layers (Layers 2–41) stand for  $12 \times 12$  WT data. Each layer includes dissimilar WT

features of  $12 \times 12$  pixels except the first layer of SLA data. To discriminate the layered eddy information, we normalized the 1-layer SLA and 40-layer WT data. It is observed that each layer contains different eddy features and the missing information becomes evident when the WT layer gradually increases to 41 (the dark blue [-1] represents the missing values). Then this new dataset will be sent to our CSA-Encoder for data compression and feature detection.

### B. CSA-Encoder Design

As shown in Fig. 5, we design a channel-spatial attention-based convolutional autoencoder (CSA-encoder), to execute data compression, denoising, and feature detection. In our experiment, it is constituted by an encoder and decoder, with channel and spatial attention modules in front of the encoder. The attention modules weight original data and output attended data to the encoder for data compression. The encoder then exports compressed data which are flattened and supplemented to the input data of TA-GRU. And the decoder finally expands compressed data to output restored data, taking the original data as the target.

Based on convolutional block attention module (CBAM) [54], two improved convolutional attention modules, such as channel and spatial attention modules are designed to refine eddy spatial characteristics. And CBAM calculates channel and spatial weights with both maxpooling and avgpooling layers. Because CBAM unsupervisedly recognizes image data in the range from 0 to 255, it does not demand a minpooling layer which always outputs 0. While the minimum values of our normalized eddy data are negative instead of 0, which prompts us to add minpooling layers to two modules for identifying characteristics from negative values automatically.

Fig. 6 shows the design of channel attention module. In this module,  $41 \times 12 \times 12$  input data are sent to global maxpooling, global avgpooling, and global minpooling layers, for aggregating and compressing information of maximum, average, and minimum values separately. Two convolutional layers (Conv1 and Conv2 in Fig. 6) are set to process the aggregated data and generate three corresponding 1-D weight vectors:  $W_{\max}$ ,  $W_{\text{avg}}$ , and  $W_{\min}$  in Fig. 6. The sum of three vectors is processed by sigmoid function  $\sigma$  to gain channel attention weight vector  $\mathbf{W}_c$ , which pays different attention to channels of input data. The kernel sizes of Conv1 and Conv2 in Fig. 6 are set to  $1 \times 1$  by (5) for making the channel attention weights' dimension match the combined data of SLA and WT. Furthermore, (4) executes summation and realizes sigmoid function on three weight vectors updated by global pooling and convolutional layers, to integrate three pooled data to the vector  $\mathbf{W}_c$ . Then multiplied by each channel of original data by (6), the channel attention weight  $\mathbf{W}_c$  makes convolutional neural networks focus on valuable channels and ignore unimportant channels. In (4), MaxPooling, AvgPooling, and MinPooling represent three global pooling layers, whose output size is  $41 \times 1 \times 1$ . Equation (5) means that  $\text{Conv}_c$  consists of two 2-D convolution layers (Conv1 and Conv2) and sigmoid function. The dimension of Conv1 is calculated by  $\frac{41}{r} \times 1 \times 1$ , where  $r$  (compression ratio)

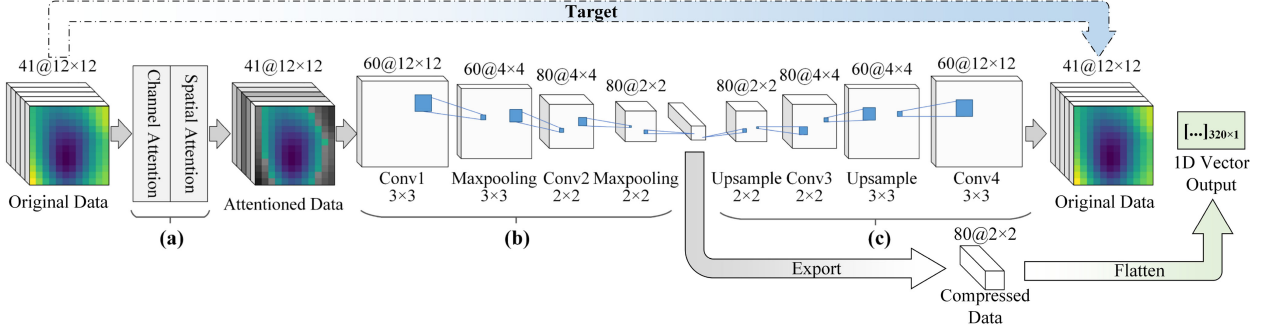


Fig. 5. Structure of CSA-encoder. (a) Convolutional attention modules calculate channel and spatial weights of original data points, and multiply both of them to obtain attended data. (b) Encoder utilizes convolutional and maxpooling layers to compress data, which is output and flatten for TA-GRU. (c) Decoder increases the dimension of compressed data and restores it to complete original data by deconvolutional and upsampling layers. (a) Attention mechanism. (b) Encoder. (c) Decoder.

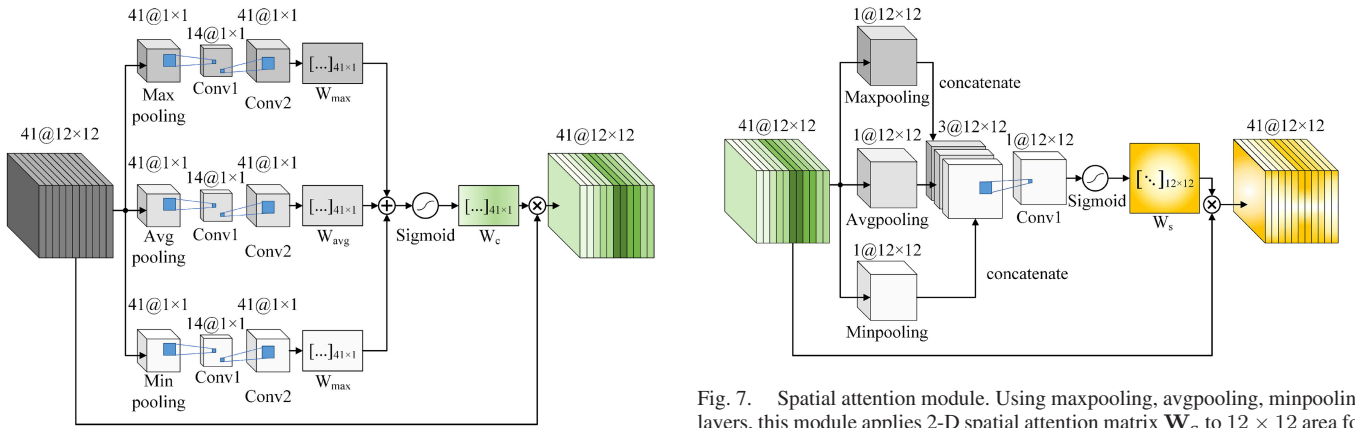


Fig. 6. Channel attention module. Calculated by maxpooling, avgpooling, minpooling layers, and convolutional layer, 1-D channel attention vector  $\mathbf{W}_c$  ensures that 41 channels can gain different attentions.

is set to 3. So the dimension of Conv1 is  $14 \times 1 \times 1$ , while Conv2 is  $41 \times 1 \times 1$  for expanding dimension. In (6),  $\mathbf{W}_c$  represents 1-D channel attention weight vector,  $\mathbf{I}_c$  is original 3-D input data, and  $\mathbf{O}_c$  means weighted data output by channel attention module

$$\mathbf{W}_c = \sigma(\text{Conv}_c(\text{MaxPooling}(\mathbf{I}_c)) + \text{Conv}_c(\text{AvgPooling}(\mathbf{I}_c)) + \text{Conv}_c(\text{MinPooling}(\mathbf{I}_c))) \quad (4)$$

$$\text{Conv}_c = \text{Conv}2^{41 \times 1 \times 1}(\sigma(\text{Conv}1^{14 \times 1 \times 1})) \quad (5)$$

$$\mathbf{O}_c = \mathbf{W}_c \odot \mathbf{I}_c. \quad (6)$$

To generate spatial attention weights for spatial attention mechanism in Fig. 5(a), the output  $\mathbf{O}_c$  of channel attention module is set to be the input  $\mathbf{I}_s$  of spatial attention module in (7). Illustrated by Fig. 7, spatial attention module also contains maxpooling, avgpooling, minpooling layers, and a convolutional layer, but pooling layers' outputs are concatenated to  $3 \times 12 \times 12$  data through (8), instead of being summarized in channel attention module. Then the concatenated data are sent to convolutional layer  $\text{Conv}_s$  (including Conv1 in Fig. 7), where configuration parameters are given by (9) for setting kernel size

Fig. 7. Spatial attention module. Using maxpooling, avgpooling, minpooling layers, this module applies 2-D spatial attention matrix  $\mathbf{W}_s$  to  $12 \times 12$  area for focusing on spatial characteristics.

to  $1 \times 1$ . After being processed by the sigmoid function, 2-D spatial attention weight matrix  $\mathbf{W}_s$  is calculated and multiplied by original data through (10), for making data points more relevant while training prediction model. We can obtain the input  $\mathbf{I}_s$  from output of channel attention module  $\mathbf{O}_c$ .  $\mathbf{W}_s$  represents 2-D spatial attention matrix, and  $\mathbf{O}_s$  is final spatial attention output. The concatenated result dimension of MaxPooling, AvgPooling, and MinPooling layers is  $3 \times 12 \times 12$ , and Conv<sub>s</sub> layer's dimension is  $1 \times 12 \times 12$ . After 2-D convolutional layer and sigmoid function conduct data compression in the size of  $12 \times 12$ ,  $\mathbf{W}_s$  applies spatial weight matrix to all data points in  $\mathbf{I}_s$

$$\mathbf{I}_s = \mathbf{O}_c \quad (7)$$

$$\mathbf{W}_s = \sigma(\text{Conv}_s([\text{MaxPooling}(\mathbf{I}_s) + \text{AvgPooling}(\mathbf{I}_s) + \text{MinPooling}(\mathbf{I}_s)])) \quad (8)$$

$$\text{Conv}_s = \text{Conv}1^{1 \times 12 \times 12} \quad (9)$$

$$\mathbf{O}_s = \mathbf{W}_s \odot \mathbf{I}_s. \quad (10)$$

In CSA-encoder, the encoder component is displaced in Fig. 5(b), which contains two convolutional layers (Conv1 and Conv2 in Fig. 5) and two maxpooling layers with tanh activation function. There are 60 kernels in first convolutional layer and



TABLE IV  
STRUCTURE AND PARAMETERS OF ENCODER

Layer	Kernel size	Padding	Stride	Output size
Input	-	-	-	41×12×12
Conv	60×3×3	1	1	60×12×12
Tanh	-	-	-	-
Maxpooling	3×3	-	-	60×4×4
Conv	80×2×2	1	1	80×4×4
Tanh	-	-	-	-
Maxpooling	2×2	-	-	80×2×2
Output	-	-	-	80×2×2

TABLE V  
STRUCTURE AND PARAMETERS OF DECODER

Layer	Kernel size	Padding	Stride	Dilation	Output size
Input	-	-	-	-	80×2×2
Upsample	2×2	-	-	-	80×2×2
Deconv	60×2×2	1	1	2	80×4×4
Tanh	-	-	-	-	-
Upsample	3×3	-	-	-	60×12×12
Deconv	41×3×3	1	1	1	41×12×12
Tanh	-	-	-	-	-
Output	-	-	-	-	41×12×12

80 kernels in the second layer, to execute dimension reduction twice for efficiently compressing eddy data. The refined data with the size of 80×2×2 can be sent to our decoder in Fig. 5(c). Table IV displays encoder's configuration in detail. Contrary to encoder, the decoder in Fig. 5(c) has two deconvolutional layers (Conv3 and Conv4 in Fig. 5) and two upsampling layers. The first deconvolutional layer (Conv3) consists of 80 kernels and the second layer (Conv4) owns 60 different kernels, to achieve dimensional expansion and interpolation for data restoration. Decoder's properties are listed in Table V.

### C. TA-GRU Design

The time sequence prediction problem requires network own memory ability, such as RNN, LSTM, and GRU, which can memorize the data of previous moments to calculate the future data. But RNN has gradient explosion and extinction problems, and LSTM network has massive parameters in input gate, forget gate, and output gate. Compared with RNN and LSTM, GRU owns less parameters, as GRU cell only has two gates: update gate and reset gate, given by

$$z_t = \sigma(W^z x_t + U^z h_{t-1}) \quad (11)$$

$$r_t = \sigma(W^r x_t + U^r h_{t-1}) \quad (12)$$

where  $x_t$  is 1-D input vector at time step  $t$  and  $h_{t-1}$  means GRU hidden state at time step  $t-1$  output by last GRU cell.  $W^z$ ,  $U^z$  or  $W^r$ ,  $U^r$  are weight matrices of update gate or reset gate, which will be optimized in training. The range of  $z_t$  and  $r_t$  are between 0 and 1 after sigmoid function  $\sigma$ , to decide memorized part and forgotten part in

$$h'_t = \tanh(Wx_t + r_t \odot Uh_{t-1}) \quad (13)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \quad (14)$$

where  $W$  and  $U$  are also weight matrices,  $h'_t$  presents new candidate state, and  $h_t$  means final state of this GRU cell.  $Uh_{t-1}$  is processed by  $r_t$  to drop out the information that should be forgotten, and is added to  $Wx_t$ . Then the sum is compressed by tanh function, to calculate  $h'_t$ . By  $z_t$  and Hadamard product  $\odot$  (also as known as the elementwise product), the proportions of  $h'_t$  and  $h_{t-1}$  get adjustment, and output  $h_t$ . If in last GRU cell,  $h_t$  is the final result of the current GRU layer. Otherwise,  $h_t$  will be sent to next cell to be its input state, and repeat above calculation recurrently.

But the output of traditional GRU network only relies on hidden state of the last cell, which ignores all previous cells' states. To avoid this drawback, we design a temporal attention layer between GRU hidden layer and FC output layer. This temporal attention layer can collect hidden states at all time steps, and then creates and adjusts different temporal attention weights for different time steps. Finally, the product of states and corresponding attention weights goes through fully connected layer, producing prediction results with the same dimension of ground truth.

Fig. 8 shows the structure of GRU hidden layer and the temporal attention layer. The GRU hidden layer consists of seven (equal to the number of input days) GRU cells (Cell<sub>t-1</sub>, Cell<sub>t</sub>, Cell<sub>t+1</sub>, etc). In Fig. 8(a), the update and reset gate in Cell<sub>t</sub> balance the last state  $h_{t-1}$  and input  $x_t$  by  $\sigma$ ,  $\odot$  and tanh functions, and then calculate state  $h_t$ , which will be sent to Cell<sub>t+1</sub> for calculating the next state  $h_{t+1}$ . The cell structures of TA-GRU and traditional GRU are similar, but their processing methods of hidden states are completely different, as traditional GRU networks (as well as RNN and LSTM) only gather information from the last state  $h_n$ , while TA-GRU memorizes all states ( $h_1, h_2, \dots, h_n$ ). As shown in Fig. 8(b), the vector of all states ( $\langle h_1, h_2, \dots, h_n \rangle$ ) is entered into the temporal attention layer and calculated by (15), which uses a transition vector  $\omega^T$  to generate temporal attention weight vector  $\alpha$ . And then (16) multiples attention weights [ $\alpha_1, \alpha_2, \dots, \alpha_n$  in Fig. 8(b)] and hidden states one by one, obtaining attentioned hidden states  $\langle h'_1, h'_2, \dots, h'_n \rangle$ , to help TA-GRU focus on states in vital days

$$\alpha = \text{softmax}(\omega^T \tanh(\langle h_1, h_2, \dots, h_n \rangle)) \quad (15)$$

$$\langle h'_1, h'_2, \dots, h'_n \rangle = \tanh(\alpha \odot \langle h_1, h_2, \dots, h_n \rangle) \quad (16)$$

In (15),  $\langle h_1, h_2, \dots, h_n \rangle$  means the vector of hidden states at all time steps,  $\alpha$  is 1-D temporal attention weight vector for hidden states. The states from GRU are calculated by tanh function to convert values into  $-1$  to  $1$ .  $\omega^T$  is a transition vector, which is used to adjust  $\alpha$  dynamically by hidden states. For the product of  $\langle h_1, h_2, \dots, h_n \rangle$  and  $\omega^T$ , softmax function ensures that the summation is 1. In (16), by Hadamard product  $\odot$ , attention weights in  $\alpha$  are multiplied by original states, to obtain temporal attentioned states  $\langle h'_1, h'_2, \dots, h'_n \rangle$ .

## IV. VALIDATION RESULTS

### A. Experiment Configuration

Our experiment environment is Python 3.6 with PyTorch 1.6 deep learning library. The development platform is Windows 10



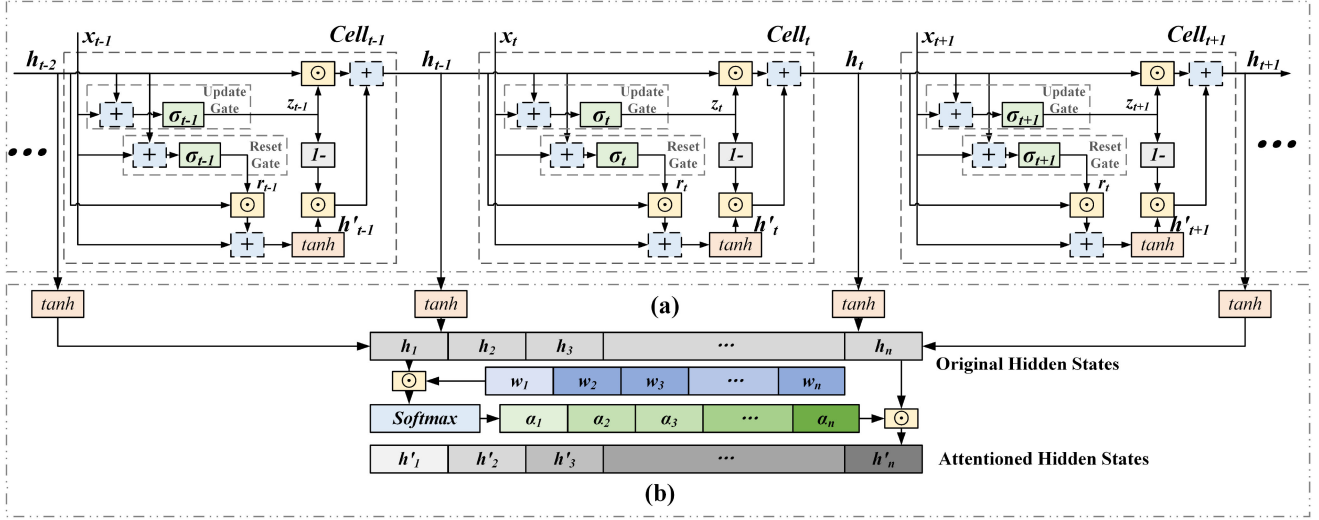


Fig. 8. Architecture of TA-GRU hidden layer and temporal attention layer. (a) Hidden layer of TA-GRU uses update and reset gates to obtain cell hidden states from input time sequences. (b) Temporal attention layer of TA-GRU gathers all states, and utilizes temporal attention weight vector  $\alpha$  to generate attentioned states. (a) GRU hidden layer. (b) Temporal attention layer.

system, and a graphical card of NVIDIA 2070 with 8 G RAM is used to train networks. Since all of three original datasets are NetCDF files, the netCDF4 library is utilized to read these data files. In data processing, NumPY library is involved, providing the data format of ndarray, which is used to store and process heterogeneous eddy data

$$\text{Ang} = \sin^2\left(\frac{\text{lat}_{\text{pred}} - \text{lat}_{\text{true}}}{2}\right) \quad (17)$$

$$+ \cos(\text{lat}_{\text{true}}) \cos(\text{lat}_{\text{pred}}) \\ \times \sin^2\left(\frac{\text{lon}_{\text{pred}} - \text{lon}_{\text{true}}}{2}\right)$$

$$\text{Dis} = 2 \times 6370 \times \arcsin(\text{Ang}). \quad (18)$$

The mean square error (mse) is selected as the loss function of the CSA-encoder and TA-GRU, to measure latitude and longitude errors between network output and ground truth. Since latitude and longitude errors represent distinct distances, (17) and (18) are imported for calculating mean distance center errors.  $\text{lat}_{\text{pred}}$  and  $\text{lat}_{\text{true}}$  indicate predicted and true latitudes as well as longitude parameters. Equation (17) can obtain the Earth's center angle Ang between predicted and true points. Then (18) utilizes Ang to get the geographic distance  $Dis$  between predicted and true locations.

And the mini-batch gradient descent (MBGD) and adaptive moment estimation (Adam) algorithms are utilized to execute mse loss reduction, whose batch size is 250. Three decreasing learning rates are implemented in different training stages: 0.0003 in epoch 0–750, 0.000045 in epoch 750–2500, and 0.000007 after epoch 2500, to accelerate the early loss decline speed and make the final loss approximate local optimal solution.

For 2718 eddy trajectories provided by TCME, we select 70%, 15%, and 15% of trajectory records as training, validation, and test sets, respectively. After feeding training datasets into our framework, 10 000 epochs can be executed if there is no overfitting problem. Our TA-GRU exploits a sliding prediction method, as the prediction result of the first day can be used as the

input for the second day's prediction. Hence the neural network can predict the trajectories of the next seven days or more in turn.

### B. Comparison of MSE Loss in CSA-Encoder

To investigate the influence between input layer number and mse loss in CSA-encoder, we compare the results for both channel and spatial attention mechanisms with different numbers of input data layers (Total Input Data Layers: 1, 6, 11, 16, 21, 31, 41) in Table VI. For 1-layer input data with an SLA layer or a WT layer, our CSA-encoder with both channel and spatial attention achieves the lowest MSE loss by 0.4453 and 0.4039. Since channel and spatial attention in CSA-encoder can selectively capture the eddy characteristics through multiple data input layers, the autoencoder gains the lowest loss among different cases. In the case of input data with 6 and 11 layers, due to more involved information, CSA-encoder gains loss of 0.2814 and 0.2667, better than other no attention-based autoencoders or the cases of one input layer. Similarly, as the layer number rises to 16, 21, 31, and 41, the mechanism with both channel and spatial attention still achieves lower losses by 0.2702, 0.2750, 0.2886, and 0.3111 than normal encoder without channel and spatial attention chosen as our baseline by 0.2810, 0.2950, 0.4032, and 0.4190, respectively. If only channel attention or spatial attention is selected, the encoder just averagely achieves 6.8% performance improvement in 1, 6, 11, 16, and 21 input layers (accordingly, 1SLA+5WT layers, 1SLA+10WT layers, 1SLA+15WT layers, and 1SLA+20WT layers) than the baseline, while our CSA-encoder can achieve 10.2% improvement than the baseline. When input WT layer number grows from 0 to 10 (from 0 to 25 m), the mse loss reduces from 0.44 to 0.26 gradually, since added WT data bring more valid eddy information. However, the center axis of a eddy is listed rather than vertical perpendicular to the Earth ground, deeper WT layer will be deviated from eddy center [65], [66]. Therefore,

TABLE VI  
COMPARISON OF AUTOENCODER LOSSES WITH DIFFERENT INPUT LAYERS AND ATTENTION MODULES

SLA input layers	WT input layers	Total input layers	Channel attention	Spatial attention	MSE loss
1	0	1	✓	✓	<b>0.4453</b>
1	0	1	✓		0.4583
1	0	1		✓	0.4857
1	0	1			0.4990
0	1	1	✓	✓	<b>0.4039</b>
0	1	1	✓		0.4057
0	1	1		✓	0.4117
0	1	1			0.4294
1	5	6	✓	✓	<b>0.2814</b>
1	5	6	✓		0.2848
1	5	6		✓	0.2854
1	5	6			0.2907
1	10	11	✓	✓	<b>0.2667</b>
1	10	11	✓		0.2682
1	10	11		✓	0.2707
1	10	11			0.2728
1	15	16	✓	✓	<b>0.2702</b>
1	15	16	✓		0.2761
1	15	16		✓	0.2776
1	15	16			0.2810
1	20	21	✓	✓	<b>0.2750</b>
1	20	21	✓		0.2820
1	20	21		✓	0.2905
1	20	21			0.2950
1	30	31	✓	✓	<b>0.2886</b>
1	30	31	✓		0.2975
1	30	31		✓	0.3105
1	30	31			0.4032
1	40	41	✓	✓	<b>0.3111</b>
1	40	41	✓		0.3557
1	40	41		✓	0.3783
1	40	41			0.4190

11 input layers with both attention mechanisms gain the lowest MSE loss of 0.2667. The bold texts mean the lowest values of MSE loss, compared with the adjacent results with different configurations between two horizontal lines.

deeper layers only provide a part of the eddy information and the signal-to-noise ratio may also be increasingly worse. So, contrary to 0–10 layers, when WT layer number increases from 15 to 40 (from 50 to 5000 m), loss value rises from 0.27 to 0.31 instead of reducing, since larger distance deviation and noise data in 15–40 WT layers compromise data quality and autoencoder performance. When the input layer numbers are set to 6, 11, 16, and 21, we can obtain similar losses in detail, as shown in the following Table VII.

Furthermore, to probe the effect of compressed data layer number on CSA-encoder's mse loss, the results with different compressed layer numbers (compressed data layers: 1, 5, 10, 20, 30, 40, 50, 60) are compared in Table VII. For four input layer number cases of 6, 11, 16, and 21 (1SLA+5WT layers, 1SLA+10WT layers, 1SLA+15WT layers, and 1SLA+20WT layers), 1-layer compressed data have losses of 0.4680, 0.4549, 0.4953, and 0.5429, always larger than other compressed layer numbers between 5 and 60, since its highest compression ratio (99.54%, 99.75%, 99.83%, and 99.87%) induces majority of eddy information to be lost. For six-layer input case, with the compressed layer number increasing from 1 to 30, the mse loss

TABLE VII  
FINE-GRAINED COMPARISON OF CSA-ENCODER LOSSES WITH DIFFERENT LAYERS OF COMPRESSED DATA

SLA input layers	WT input layers	Compressed data layers	Compression ratio	MSE loss
1	5	1	99.54%	0.4680
1	5	5	97.69%	0.2924
1	5	10	95.37%	0.2680
1	5	20	90.74%	0.2641
1	5	30	86.11%	<b>0.2637</b>
1	5	40	81.48%	0.2733
1	5	50	76.85%	0.2740
1	5	60	72.22%	0.2764
1	10	1	99.75%	0.4549
1	10	5	98.74%	0.2998
1	10	10	97.47%	0.2717
1	10	20	94.95%	0.2542
1	10	30	92.42%	<b>0.2528</b>
1	10	40	89.90%	0.2531
1	10	50	87.37%	0.2611
1	10	60	84.85%	0.2649
1	15	1	99.83%	0.4953
1	15	5	99.13%	0.3448
1	15	10	98.26%	0.2904
1	15	20	96.53%	0.2668
1	15	30	94.79%	<b>0.2641</b>
1	15	40	93.06%	0.2647
1	15	50	91.32%	0.2717
1	15	60	89.58%	0.2828
1	20	1	99.87%	0.5429
1	20	5	99.34%	0.3989
1	20	10	98.68%	0.3034
1	20	20	97.35%	0.2754
1	20	30	96.03%	<b>0.2649</b>
1	20	40	94.71%	0.2667
1	20	50	93.39%	0.2735
1	20	60	92.06%	0.2777

Data with 1 SLA layer, 11 WT layers, and 30 compressed layers get the lowest MSE loss of 0.2528.

The bold texts mean the lowest values of MSE loss, compared with the adjacent results with different configurations between two horizontal lines.

keeps reducing from 0.4680 to 0.2637, since there are more related data retained. Similarly, for other input layer numbers like 11, 16, and 21, the compressed layer number growth also brings more related information, helping CSA-encoder decrease loss by 46.48% averagely. However, when compressed layer number climbs to 40, 50, and 60, the six-layer input data case gets rising loss values of 0.2733, 0.2740, 0.2764, which are 3.5%, 3.7%, and 4.5% higher than loss value of 30 compressed layers, because more layer data also introduce more redundant information. Other cases of input layer number (11, 16, and 21 layers) also get the similar rising losses when compressed layer numbers are 40, 50, and 60, with about 0.35%, 3.1%, and 5.5% loss increasing, compared to 30 compressed layers. The lowest loss value, 0.2528, appears in the case of 11 input layers (1SLA+10WT layers) and 30 compressed layers, contributed by more input characteristics and less redundant compressed data. So we select 30-layer compressed data encoded by 1 SLA layer and 11 WT layers, for data compression of CSA-encoder and subsequent trajectory prediction of TA-GRU.

Fig. 9 further examines mse losses of decoded eddy data with different numbers of learning epochs. The subfigures of

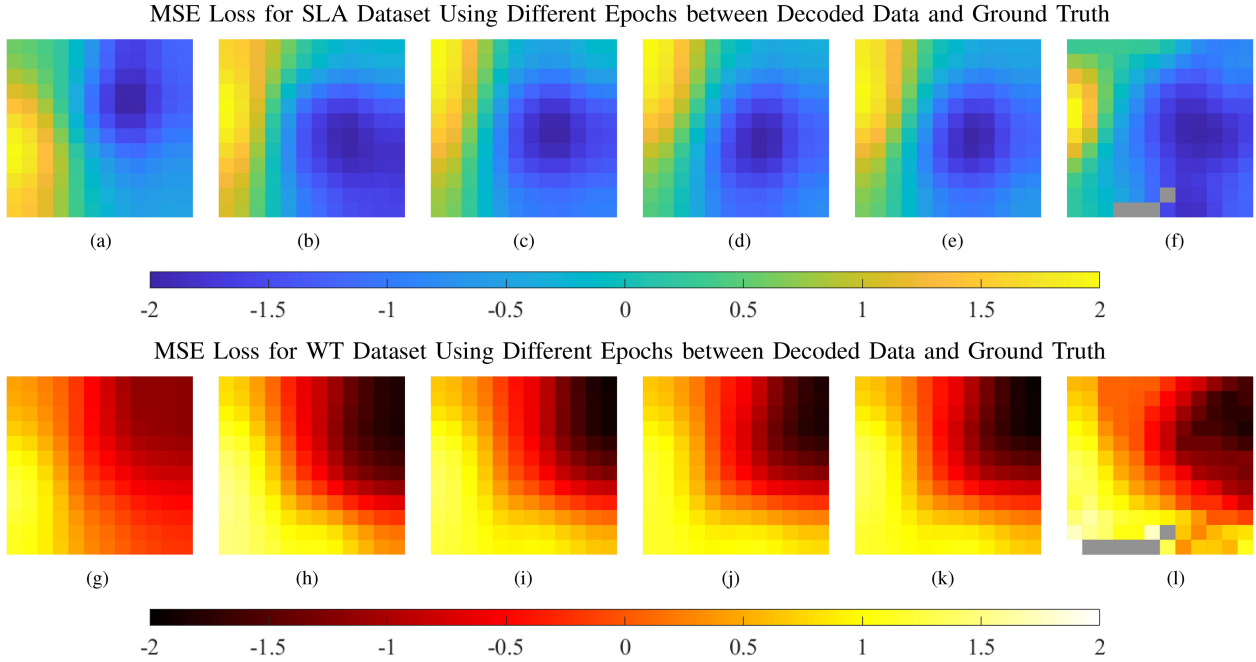


Fig. 9. Comparison of MSE loss between decoded data and ground truth in data range:  $[-2, 2]$ . Gray area in ground truth are for uncollected or missing eddy data points. In epoch 0-5000, the MSE losses of SLA and WT decrease from 0.87, 0.95 to 0.23, 0.26, respectively. (a) Loss:0.87, Epo. 0. (b) Loss:0.36, Epo. 500. (c) Loss:0.26, Epo. 1000. (d) Loss:0.25, Epo. 2000. (e) Loss:0.23, Epo. 5000. (f) Ground Truth. (g) Loss:0.95, Epo. 0. (h) Loss:0.38, Epo. 500. (i) Loss:0.29, Epo. 1000. (j) Loss:0.27, Epo. 2000. (k) Loss:0.26, Epo. 5000. (l) Ground Truth.

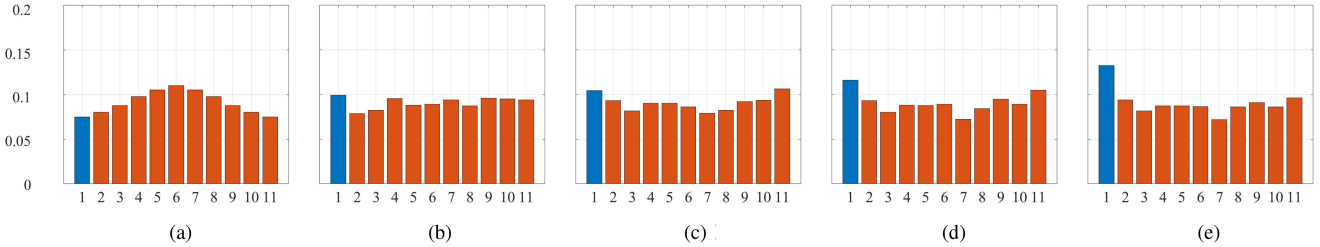


Fig. 10. Channel attention weights variation from epoch 0 to 5000. 11 bars mean different attention weight values for 11 input channels, where blue bar (1) means attention weight for SLA data, and orange bars (2–11) represent ten weights for WT data. The sum of 11 weights is always 1 during a period of learning eddy properties. (a) Epoch 0. (b) Epoch 500. (c) Epoch 1000. (d) Epoch 2000. (e) Epoch 5000.

Fig. 9(a)–(f) demonstrate SLA losses and the ground truth. In this experiment, our attention mechanisms reduce SLA loss from 0.87 to 0.23 when the epoch number increases from 0 to 1000. Likewise, for decoded WT data of mesoscale eddies, the mse loss is reduced from 0.95 to 0.26 with the increasing epochs in Fig. 9(g)–(k). In the ground truth of Fig. 9(f) and (l), the gray color points stand for the missing eddy data. We find that our CSA-encoder can generate new data points to compensate the gray area of Fig. 9(f) and (i) in Fig. 9(c)–(e) and (i)–(k). It is obvious that more iterative training procedures can gradually reduce the mse loss from 0.26 to 0.23 with the epochs increasing from 1000 to 5000.

To explore how channel attention mechanism works well in training procedure, Fig. 10 displays 11 attention weights' variation for 11 input layers (1 SLA layer and 10 WT layers) in epochs 0, 500, 1000, 2000, and 5000. Although each weight keeps varying in training, the summation of 11 weights is always set to 1 by softmax function. In epoch 0 [Fig. 10(a)], the initial

weights follow a normal distribution, so the sixth layer of the combined data gets highest weight of 0.11, the first and last layers gain lowest weights of 0.075. However, from epoch 0 to epoch 5000, the weight of SLA data (blue bar) gradually increases from 0.075 to 0.133, while 10 weights of other WT layers (10 orange bars) tend to decrease to similar values (about 0.09) in epoch 5000. Shown in Fig. 10(e), SLA gets highest channel attention weight of 0.133 in epoch 5000, which is 47.7% higher than WT layers,' since SLA contains fewer missing data points and more intuitive information, beneficial to loss minimization.

### C. Prediction Result of TA-GRU

In order to examine the working process of temporal attention mechanism, Fig. 11 illustrates the variation of the temporal attention weights for seven-day input data in training, where seven bars represent seven attention weight values for seven



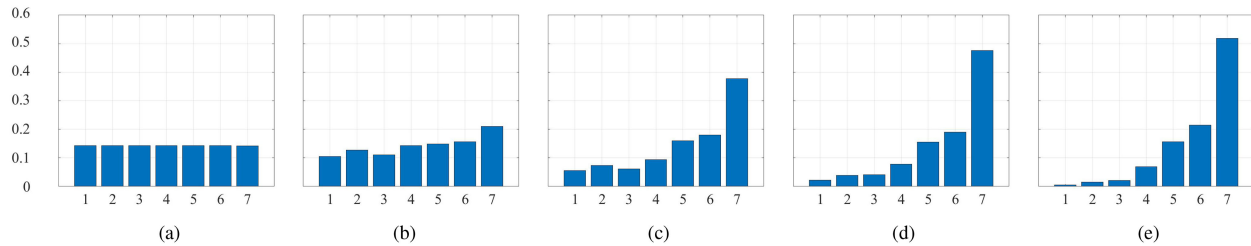


Fig. 11. Seven temporal attention weights variation from epoch 0 to 5000. Seven bars represent attention weights for seven-day input data, respectively. The temporal attention is allocated to seven weights dynamically in training, whose sum value is always 1. (a) Epoch 0. (b) Epoch 500. (c) Epoch 1000. (d) Epoch 2000. (e) Epoch 5000.

TABLE VIII  
1-DAY PREDICTION RESULTS WITH DIFFERENT DATASETS AND ATTENTION MODULES

Datasets and attention mechanisms	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11
TCME dataset	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SLA dataset	-	✓	-	✓	✓	✓	✓	✓	✓	✓	✓
WT dataset	-	-	✓	✓	✓	✓	✓	✓	✓	✓	✓
Channel attention	-	-	-	-	✓	-	-	✓	✓	-	✓
Spatial attention	-	-	-	-	-	✓	-	✓	-	✓	✓
Temporal attention	-	-	-	-	-	-	✓	-	✓	✓	✓
MSE loss value	0.0045	0.0021	0.0025	0.0015	0.0015	0.0014	0.0013	0.0013	0.0012	0.0012	<b>0.0009</b>
Center error (km)	19.9901	17.1898	17.7405	16.5187	14.3752	14.5707	13.9877	12.4937	11.8735	12.0362	<b>9.0812</b>

Case 11 achieves the lowest MSE loss of 0.0009 and smallest center distance error of 9.0812 Km, outperforming cases 1–10, whose advantage comes from the three listed attention mechanisms and the heterogeneous data integration.

The two bold entities show the lowest MSE loss and center error among the 16 cases respectively.

input days. The  $X$ -axis of Fig. 11 denotes the input weights of temporal attention for seven days, which approximately equals to 0.14 because the summation of temporal attention of 7 input channels is 1 (each weight is  $1/7$ ). In Fig. 11(a), the initial temporal attention weights are allocated to seven days with an average value of 0.1429. Fig. 11(b)–(e) show that 1–4 days' weights drop by 77.28% from epoch 500 to 5000. On the contrary, the weights of 5–7 days increase by 100.02% averagely. After training 5000 epochs in Fig. 11(e), weights of 1–4 days are reduced continuously to 0.0051, 0.0155, 0.0206, and 0.0691, while the data of 5–7 days gain higher weights, which are 0.1561, 0.2148, and 0.5188, respectively. Fig. 11(e) reveals that TA-GRU assigns more than 94% attention to the eddy data in 5–7 days. Hence TA-GRU can pay more vital attention to the data in more recent days which supply more real-time properties instead of outdated information of previous days.

Table VIII is applied to verify the effects of different datasets and attention mechanisms on prediction performance, where one-day forecasting errors in 11 different cases are compared. For Case 1 in Table VIII, TA-GRU with only TCME features gains the highest mse loss of 0.0045 and largest distance error of 19.99 km, since 1-D features in TCME contain less eddy properties. The results of Cases 2, 3, and 4 prove that the addition of SLA and WT indeed reduces distance error to 17.1898 km, and the loss of two datasets is lower than one dataset alone (with approximately 5.4% reduction). Cases 5–7 select the same datasets (TCME, SLA, and WT) and the different methods with one attention mechanism. The experimental results show that spatial, channel and temporal attention mechanisms reduce distance errors to 14.3752, 14.5707, and 13.9877 km, respectively,

and the method with all the three attention mechanisms achieves 12% error reduction lower than other methods. In Cases 8–10, different schemes with two attention mechanisms are involved for same three datasets, demonstrating further error reduction (15.2% on average) compared to a single attention mechanism applied in Cases 5–7. Case 11 preserves the best prediction performance with three datasets (TCME, SLA, and WT) and three attention mechanisms (channel, spatial, and temporal attention), by 0.0009 mse loss and 9.0817 km distance error for next one day, which is 54% lower than the baseline GRU without extra datasets or attention mechanisms.

Table IX shows the distance error results of TA-GRU and traditional RNN, LSTM, GRU, which make sliding predictions of the eddy center positions for 1–7 days, 10th day, and 14th day. For TCME dataset in Table IX, our TA-GRU achieves the lowest distance errors for seven-day continuous prediction, 10th and 14th-day forecasting. Compared with the networks with TCME dataset, such those networks with TCME, SLA, and WT datasets incur 17% degradation of distance errors on average, indicating SLA and WT datasets improve the prediction performance. In the last row of Table IX, we can observe that TA-GRU with all three datasets obtains the lowest loss and distance error for 1–14 days, with one-day error of 9.08 km, seven-day error of 23.86 km, and 14-day error of 59.06 km. In contrast with traditional RNN, LSTM, and GRU, TA-GRU can bring 65.6%, 54.9%, and 54.1% distance error reduction, respectively, as attention mechanisms help TA-GRU identify eddy's spatio-temporal characteristics more accurately.

Five forecasting curves of longitude and latitude are illustrated in Fig. 12(a) and (b) separately, reflecting four seven-day

TABLE IX  
CENTER DISTANCE ERROR COMPARISON WITH DIFFERENT NEURAL NETWORKS AND MESOSCALE EDDY DATASETS (KM)

Methods	Days in future								
	1	2	3	4	5	6	7	10	14
RNN & TCME	26.3882	29.0234	35.3642	39.3452	43.8010	49.0268	54.4620	75.6201	89.2584
LSTM & TCME	20.1182	22.6875	25.0668	29.2354	32.3845	36.8801	39.3326	60.1022	81.3175
GRU & TCME	19.7901	21.6340	24.0357	28.0654	31.1824	35.3932	37.8563	58.235	80.6201
TA-GRU & TCME	14.7704	16.0462	17.8659	19.6032	22.0368	26.1035	29.8659	41.3258	65.3504
RNN & TCME & SLA & WT	23.8654	26.8342	30.6634	35.8352	39.3321	45.0395	51.2385	70.0854	85.9053
LSTM & TCME & SLA & WT	16.7823	18.9654	20.0135	22.1325	25.8651	29.3201	35.8932	53.0230	67.0385
GRU & TCME & SLA & WT	15.5187	17.6428	19.1913	21.0632	24.3594	28.5690	34.9965	51.3001	69.3825
TA-GRU & TCME & SLA & WT	<b>9.0812</b>	<b>11.3424</b>	<b>12.9354</b>	<b>14.6387</b>	<b>17.9652</b>	<b>20.0354</b>	<b>23.8624</b>	<b>35.5384</b>	<b>59.0615</b>

TA-GRU with temporal attention mechanism achieves the best performance in three eddy datasets TCME, SLA, and WT for seven-day continuous prediction, 10th day and 14th day nowcasting in future.

These bold values are the smallest center distance errors in 1-14 days, which are compared with results of the same column.

TABLE X  
COMPARISON OF COMPUTATIONAL TIME FOR DIFFERENT NEURAL NETWORKS

Networks	Experiment number										Average Time (ms)
	1	2	3	4	5	6	7	8	9	10	
RNN	1.905	2.500	2.004	2.003	2.012	2.157	2.015	2.501	2.000	1.971	<b>2.107</b>
LSTM	2.046	2.026	2.021	1.998	3.015	3.207	3.018	2.483	2.014	2.015	<b>2.384</b>
GRU	1.984	2.015	1.999	1.986	2.000	1.999	1.986	2.032	2.500	2.001	<b>2.050</b>
TA-GRU	2.500	2.000	2.519	2.018	2.000	2.504	2.018	2.287	2.501	2.007	<b>2.235</b>

<sup>3</sup>GRU is the fastest with 2.050 Ms, TA-GRU is 0.18 ms longer than GRU, since temporal attention layer brings extra parameters, such as  $\alpha$  and  $\omega^T$ , and LSTM is the slowest with 2.384 ms as its cell structure is the most complicated among the four networks.

The bold values represent the average computational time for the first ten columns in the corresponding row.

prediction results of TA-GRU, GRU, LSTM, RNN, and the ground truth of the eddy trajectory. It is obvious that the red curve of TA-GRU is much closer to the ground truth (black curve) than GRU, LSTM, and RNN, since TA-GRU iteratively captures more eddy properties. Compared to GRU, LSTM, and RNN, TA-GRU has 38%, 45%, and 65% distance error declines, respectively. In Fig. 12(b), for the inflection point ( $9.13^\circ N$ ) of ground truth's latitude on the third day, TA-GRU's inflection point ( $9.14^\circ N$ ) appears approximately on the fourth day, while the inflection points ( $9.15^\circ N$ ,  $9.14^\circ N$ ,  $9.13^\circ N$ ) of GRU, LSTM, and RNN occur on the fifth day. Hence TA-GRU leads to the lower latency for longtime trajectory prediction.

#### D. Computational Time Result of TA-GRU

Table X shows the the computational time results of baseline RNN, LSTM, GRU, and TA-GRU, which repeats prediction experiments ten times and achieves the average time costs. In the four networks, GRU's time is the lowest with 2.050 ms, since there are less parameters ( $z_t, r_t, W, U, h'_t$ , etc.) in its simplified two-gate cell structure. And RNN preserves the second least time consumption of 2.107 ms, which is 2.7% longer than GRU. TA-GRU achieves the computational time of 2.235 ms, which is 0.18-ms longer than baseline GRU. As TA-GRU adds temporal attention layer after the GRU hidden layer, there are more parameter vectors (such as  $\alpha$ ,  $\omega^T$ , and  $h'$ , which causes approximately 0.6-KB extra storage space in memory), which slightly increases the computational complexity by approximately 9% time prolongation. Obviously, LSTM leads to the largest time cost by 2.384 ms, which is 16% longer than GRU

and 7% longer than TA-GRU, as LSTM has the most parameters in every operation cell of three gates.

In comparison, the new temporal attention mechanism increases the computation time of TA-GRU by 0.18 ms over baseline GRU and 0.13 ms over RNN, but it also reduces time by 7% (0.15 ms) compared to LSTM with more complex cell structure. In addition, although our TA-GRU causes more computational efforts than GRU and RNN, the computational overhead is negligible as TA-GRU gains approximately 65%, 55%, and 54% distance error reduction over traditional RNN, LSTM, and GRU.

#### V. CONCLUSION

In this article, utilizing multidimensional datasets covering eddy stereoscopic structure, a spatio-temporal attention-based deep learning framework is proposed to accurately predict trajectories of mesoscale eddies in the SCS, which consists of three parts: 1) data construction and processing, 2) data compression by CSA-encoder, and 3) forecasting network TA-GRU. The heterogeneous datasets in the SCS are extracted, combined, and processed first. To capture critical spatial features in heterogeneous eddy data, CSA-encoder is proposed to deduct redundant data, denoise, and compress data. The channel and spatial attention mechanisms in CSA-encoder are dedicated to vertical and horizontal space feature detection, and temporal attention mechanism in TA-GRU contributes to capturing key time sequence characteristics, which reduces the mse losses significantly. Our cross-validation results show that our forecasting framework with spatio-temporal attention mechanisms achieves the lowest forecasting error whose range is from 9 to

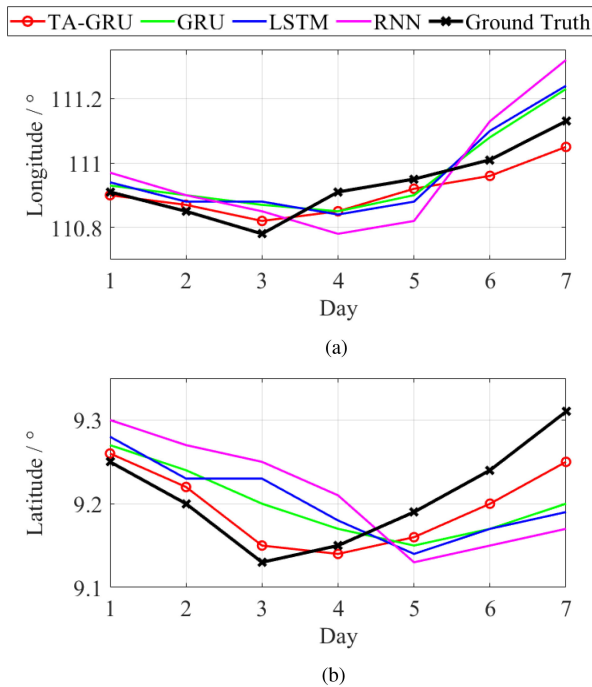


Fig. 12. Predicted results through four different neural networks in future seven days. The black line is ground truth and the red represents our predicted latitude and longitude by TA-GRU. The predicted results of TA-GRU are the closest to the true eddy trajectories. (a) 7-Day Longitude Prediction. (b) 7-Day Latitude Prediction.

23 km in seven-day prediction, and provides 57%, 44%, and 42% prediction performance improvements compared to traditional RNN, GRU, and LSTM with single TCME. CSA-encoder and TA-GRU can also be applied to eddy trajectory prediction in other ocean regions, such as the East China Sea and the North Indian Ocean.

The improved accuracy proves the significant advantages of our framework over previous eddy predicting networks, which provides the strategies applying heterogeneous data and spatio-temporal attention mechanisms for other prediction of oceanic phenomena which own stereoscopic structure and time-series data. Meanwhile, in the future, it is worth further exploring the impact of more related datasets on the performance of eddy trajectory prediction, such as geostrophic flow velocity, ocean salinity, wind farm, and submarine topography.

## REFERENCES

- [1] D. B. Chelton, M. G. Schlax, and R. M. Samelson, "Global observations of nonlinear mesoscale eddies," *Prog. Oceanogr.*, vol. 91, no. 2, pp. 167–216, 2011.
- [2] Y. Zhang, Z. Liu, Y. Zhao, W. Wang, J. Li, and J. Xu, "Mesoscale eddies transport deep-sea sediments," *Sci. Rep.*, vol. 4, no. 1, 2014, Art. no. 5937.
- [3] Z. Zhang, W. Wang, and B. Qiu, "Oceanic mass transport by mesoscale eddies," *Science*, vol. 345, no. 6194, pp. 322–324, 2014.
- [4] L. Xu, P. Li, S.-P. Xie, Q. Liu, C. Liu, and W. Gao, "Observing mesoscale eddy effects on mode-water subduction and transport in the North Pacific," *Nature Commun.*, vol. 7, no. 1, pp. 1–9, 2016.
- [5] L. Xu, S. Xie, J. L. McClean, Q. Liu, and H. Sasaki, "Mesoscale eddy effects on the subduction of north pacific mode waters," *J. Geophysical Res., Oceans*, vol. 119, no. 8, pp. 4867–4886, 2014.
- [6] C. Eden and H. Dietze, "Effects of mesoscale eddy/wind interactions on biological new production and eddy kinetic energy," *J. Geophysical Res. Oceans*, vol. 114, no. C5, 2009, Art. no. C05023.
- [7] D. Tyler, J. R. Barnes, and E. D. Skillingstad, "Mesoscale and large-eddy simulation model studies of the martian atmosphere in support of phoenix," *J. Geophysical Res.: Planets*, vol. 113, no. E3, 2008, Art. no. E00A12.
- [8] C. Horvat, E. Tziperman, and J.-M. Campin, "Interaction of sea ice floe size, ocean eddies, and sea ice melting," *Geophysical Res. Lett.*, vol. 43, no. 15, pp. 8083–8090, 2016.
- [9] Z. Fan, G. Zhong, H. Wei, and H. Li, "EDNET: A mesoscale eddy detection network with multi-modal data," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.
- [10] A. Rubio, B. Blanke, S. Speich, N. Grima, and C. Roy, "Mesoscale eddy activity in the southern Benguela upwelling system from satellite altimetry and model data," *Prog. Oceanogr.*, vol. 83, no. 1/4, pp. 288–295, 2009.
- [11] W. Alpers *et al.*, "A small-scale oceanic eddy off the coast of West Africa studied by multi-sensor satellite and surface drifter data," *Remote Sens. Environ.*, vol. 129, pp. 132–143, 2013.
- [12] A. R. Robinson *et al.*, "Real-time dynamical forecast of ocean synoptic/mesoscale eddies," *Nature*, vol. 309, no. 5971, pp. 781–783, 1984.
- [13] H. E. Hurlburt, "The potential for ocean prediction and the role of altimeter data," *Mar. Geodesy*, vol. 8, no. 1/4, pp. 17–66, 1984.
- [14] A. R. Robinson and W. G. Leslie, "Estimation and prediction of oceanic eddy fields," *Prog. Oceanogr.*, vol. 14, pp. 485–510, 1985.
- [15] J. Wang, "A. nowcast/forecast system for coastal ocean circulation using simple nudging data assimilation," *J. Atmospheric Ocean. Technol.*, vol. 18, no. 6, pp. 1037–1047, 2001.
- [16] A. Klocker and R. Abernathy, "Global patterns of mesoscale eddy properties and diffusivities," *J. Phys. Oceanogr.*, vol. 44, no. 3, pp. 1030–1046, 2014.
- [17] J. Faghmous *et al.*, "Multiple hypothesis object tracking for unsupervised self-learning: An ocean eddy tracking application," in *Proc. AAAI Conf. Artif. Intell.*, 27, 2013, pp. 1227–1283.
- [18] K. Franz, R. Roscher, A. Milioto, S. Wenzel, and J. Kusche, "Ocean eddy identification and tracking using neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6887–6890.
- [19] X. Sun, M. Zhang, J. Dong, R. Lguensat, Y. Yang, and X. Lu, "A deep framework for eddy detection and tracking from satellite sea surface height data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7224–7234, 2021.
- [20] G. Xu *et al.*, "Oceanic eddy identification using an ai scheme," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1349.
- [21] D. Huang, Y. Du, Q. He, W. Song, and A. Liotta, "DeepEddy: A simple deep architecture for mesoscale oceanic eddy detection in SAR images," in *Proc. IEEE 14th Int. Conf. Netw. Sensing Control*, pp. 673–678, 2017.
- [22] Z. Duo, W. Wang, and H. Wang, "Oceanic mesoscale eddy detection method based on deep learning," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1921.
- [23] M. D. Ashkezari, C. N. Hill, C. N. Follett, G. Forget, and M. J. Follows, "Oceanic eddy detection and lifetime forecast using machine learning methods," *Geophysical Res. Lett.*, vol. 43, no. 23, pp. 12–234, 2016.
- [24] H. E. Hurlburt *et al.*, "Eddy-resolving global ocean prediction," Naval Research Lab Stennis Space Center Ms Oceanography Div, John C. Stennis Space Center, MS, USA, Tech. Rep. ADA502848, 2009.
- [25] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.
- [26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [27] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
- [28] X. Wang, H. Wang, D. Liu, and W. Wang, "The prediction of oceanic mesoscale eddy properties and propagation trajectories based on machine learning," *Water*, vol. 12, no. 9, 2020, Art. no. 2521.
- [29] C. Ma, S. Li, A. Wang, J. Yang, and G. Chen, "Altimeter observation-based eddy nowcasting using an improved CONV-LSTM network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 783.
- [30] Y. Ren, X. Li, X. Yang, and H. Xu, "Development of a dual-attention U-Net model for SEA ice and open water classification on SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4010205.
- [31] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sens.*, vol. 10 no. 10, 2018, Art. no. 1602.



- [32] M. Yang, K. Hu, C. Li, and Z. Wei, "UW-NET: An inception-attention network for underwater image classification," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–10.
- [33] J. Liu, J. Yan, L. Wang, L. Huang, H. He, and H. Liu, "Remote sensing time series classification based on self-attention mechanism and time sequence enhancement," *Remote Sens.*, vol. 13, no. 9, 2021, Art. no. 1804.
- [34] B. Qiao, Z. Wu, Z. Tang, and G. Wu, "Sea surface temperature prediction approach based on 3D CNN and LSTM with attention mechanism," in *Proc. 23rd Int. Conf. Adv. Commun. Technol.*, 2021, pp. 342–347.
- [35] A. Immas, N. Do, and M.-R. Alam, "Real-time in situ prediction of ocean currents," *Ocean Eng.*, vol. 228, 2021, Art. no. 108922.
- [36] J. Liu, B. Jin, L. Wang, and L. Xu, "Sea surface height prediction with deep learning based on attention mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 1501605.
- [37] T. Song, J. Jiang, W. Li, and D. Xu, "A deep learning method with merged LSTM neural networks for SSHA prediction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2853–2860, 2020.
- [38] S. Chen and Y. Liu, "Migration learning based on computer vision and its application in ocean image processing," *J. Coastal Res.*, vol. 104, no. SI, pp. 281–285, 2020.
- [39] S. Fan, J. Fei, X. Guo, C. O. Yang, and A. J. Revell, "CNN LSTM accelerated turbulent flow simulation with link-wise artificial compressibility method," in *Proc. 50th Int. Conf. Parallel Process.*, 2021, pp. 1–10.
- [40] H. Y. Guo, M. X. Chen, and L. Han, "Evaluation of the Conv-GRU deep learning method for convective weather nowcasting," in *Proc. 19th Conf. Artif. Intell. Environ. Sci.*, 2020.
- [41] 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 Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 740–744, May 2020.
- [42] T. Song, Z. Wang, P. Xie, N. Han, J. Jiang, and D. Xu, "A novel dual path gated recurrent unit model for sea surface salinity prediction," *J. Atmospheric Ocean. Technol.*, vol. 37, no. 2, pp. 317–325, 2020.
- [43] R. W. Liu, M. Liang, J. Nie, W. Y. B. Lim, Y. Zhang, and M. Guizani, "Deep learning-powered vessel trajectory prediction for improving smart traffic services in maritime Internet of Things," *IEEE Trans. Netw. Sci. Eng.*, early access, Jan. 7, 2022, doi: [10.1109/TNSE.2022.3140529](https://doi.org/10.1109/TNSE.2022.3140529).
- [44] M. M. Kordmahalleh, M. G. Sefidmazgi, and A. Homaifar, "A sparse recurrent neural network for trajectory prediction of Atlantic hurricanes," in *Proc. Genetic Evol. Comput. Conf.*, 2016, pp. 957–964.
- [45] Z. I. Petrou and Y. Tian, "Prediction of sea ice motion with convolutional long short-term memory networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6865–6876, Sep. 2019.
- [46] M. D. Grossi, M. Kubat, and T. M. Özgökmen, "Predicting particle trajectories in oceanic flows using artificial neural networks," *Ocean Modelling*, vol. 156, 2020, Art. no. 101707.
- [47] Y. Yang, J. Dong, X. Sun, E. Lima, Q. Mu, and X. Wang, "A FCFC-LSTM model for sea surface temperature prediction," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 207–211, Feb. 2018.
- [48] S. Fan, N. Xiao, and S. Dong, "A novel model to predict significant wave height based on long short-term memory network," *Ocean Eng.*, vol. 205, 2020, Art. no. 107298.
- [49] F. Enigo VS *et al.*, "Forecasting significant wave height using RNN-LSTM models," in *Proc. 4th Int. Conf. Intell. Comput. Control Syst.*, 2020, pp. 1141–1146.
- [50] Z. Zhang, W. Zhao, J. Tian, and X. Liang, "A mesoscale eddy pair southwest of Taiwan and its influence on deep circulation," *J. Geophysical Res., Oceans*, vol. 118, no. 12, pp. 6479–6494, 2013.
- [51] Z. Zhang, Y. Zhang, W. Wang, and R. X. Huang, "Universal structure of mesoscale eddies in the ocean," *Geophysical Res. Lett.*, vol. 40, no. 14, pp. 3677–3681, 2013.
- [52] J. T. Coull, "fMRI studies of temporal attention: Allocating attention within, or towards, time," *Cogn. Brain Res.*, vol. 21, no. 2, pp. 216–226, 2004.
- [53] A. Martínez *et al.*, "Objects are highlighted by spatial attention," *J. Cogn. Neurosci.*, vol. 18, no. 2, pp. 298–310, 2006.
- [54] S. Woo, J. Park J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 3–19.
- [55] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 95–104.
- [56] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 4836–4845.
- [57] W. Cai, Z. Wei, R. Liu, Y. Zhuang, Y. Wang, and X. Ning, "Remote sensing image recognition based on multi-attention residual fusion networks," *ASP Trans. Pattern Recognit. Intell. Syst.*, vol. 1, no. 1, 2021.
- [58] M. D. Karl and R. Lukas, "The hawaii ocean time-series (HOT) program: Background, rationale and field implementation," *Deep Sea Research II, Topical Studies Oceanogr.*, vol. 43, no. 2/3, pp. 129–156, 1996.
- [59] K. L. Denman and A. E. Gargett, "Time and space scales of vertical mixing and advection of phytoplankton in the upper ocean," *Limnology Oceanogr.*, vol. 28, no. 5, pp. 801–815, 1983.
- [60] Y. Feng, T. Sun, and C. Li, "Study on the influence of attention mechanism in large-scale sea surface temperature prediction based on temporal convolutional network," in *Proc. Int. Conf. Mobile Multimedia Commun.*, 2021, pp. 727–735.
- [61] N. Thongniran, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathien, and P. Vateekul, "Combining attentional CNN and GRU networks for ocean current prediction based on HF radar observations," in *Proc. 8th Int. Conf. Comput. Pattern Recognit.*, 2019, pp. 440–446.
- [62] T. Song, Y. Li, F. Meng, P. Xie, and D. Xu, "A novel deep learning model by BIGRU with attention mechanism for tropical cyclone track prediction in northwest pacific," *J. Appl. Meteorol. Climatol.*, vol. 61, no. 1, pp. 3–12, 2022.
- [63] C. Xie, H. Guo, and J. Dong, "LSENet: Location and seasonality enhanced network for multi-class ocean front detection," 2021, *arXiv:2108.02455*.
- [64] Z. Zhang *et al.*, "Observed 3D structure, generation, and dissipation of oceanic mesoscale eddies in the South China Sea," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 24349.
- [65] J. Hu, J. Gan, Z. Sun, J. Zhu, and M. Dai, "Observed three-dimensional structure of a cold eddy in the southwestern South China Sea," *J. Geophysical Res. Oceans*, vol. 116, no. C5, 2011, Art. no. C05016.
- [66] J. A. Kurczyn, E. Beier, M. F. Lavín, A. Chaigneau, and V. M. Godínez, "Anatomy and evolution of a cyclonic mesoscale eddy observed in the northeastern Pacific tropical-subtropical transition zone," *J. Geophysical Res., Oceans*, vol. 118, no. 11, pp. 5931–5950, 2013.