ISPRS Journal of Photogrammetry and Remote Sensing 229 (2025) 382–395

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Sea Level Anomaly prediction with TSTA-enhanced UNet

Qinxuan Wang [a,b], Jun Bai [c,*], Yineng Li [d], Shiming Xiang [a,b], Xiaoqing Chu [e], Yue Sun [f], Tielin Zhang [g]

[a] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China
[c] Aerospace Center, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[d] State Key Laboratory of Tropical Oceanography, Key Laboratory of Science and Technology on Operational Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 511458, China
[e] State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China
[f] Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[g] Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, 200031, China

ARTICLE INFO

Keywords:
SLA prediction
UNet
Vision Transformer
Temporal-Spatial Transformer Attention

ABSTRACT

The prediction of Sea Level Anomaly (SLA) is crucial for many applications in marine and meteorological tasks. Most recently developed SLA prediction methods have been developed mainly on the framework of the Recurrent Neural Network (RNN) and its variants. These frameworks suffer from insufficient capability to capture spatial information and low computational efficiency. To address these issues, this paper proposes a novel method called UNet and Temporal-Spatial Transformer Attention (UNet-TSTA) for accurate and efficient SLA prediction. In our model, UNet serves as the backbone structure of the prediction model, enhancing the model's ability to capture features of sea surface eddies at different scales. Meanwhile, the TSTA module innovatively constructs multiple spatial–temporal planes through the free combination of temporal and spatial dimensions, utilizing the attention mechanism of the Point-by-Point Vision Transformer (P-ViT). The effective cooperation of P-ViT and CNN also enhances the training and inference speed of the model. Experimental results on real SLA datasets show that our UNet-TSTA method achieves millimeter-level average precision in predicting SLA fields for the next seven days. Compared to other advanced algorithms, our method shows significant improvements in both computational efficiency and prediction precision.

1. Introduction

Sea Level Anomaly (SLA) refers to the deviation of the ocean surface height from a long-term mean value (Bonaduce et al., 2016). The research on SLA is of great significance to many forecasting applications. Predicting SLA enables earlier and more accurate forecasts of El Niño and La Niña phenomena (Zhao et al., 2023). Additionally, SLA can assist in monitoring and predicting tsunamis and storm surges, providing early warning information before natural disasters occur, thereby reducing loss of lives and properties (Ningsih et al., 2020).

Traditional methods for predicting SLA primarily rely on physics-based mathematical models and a series of complex physical equations (Chen et al., 1998; Gregory and Lowe, 2000; Miles et al., 2014). However, physical and dynamic models may lack sufficient flexibility to handle complex ocean phenomena. Technically, it usually requires consideration of a large number of physical parameters and complex physical processes, resulting in high computational costs, which limits their application in large-scale SLA prediction.

Deep learning models can effectively learn features from large amounts of data, capture and process complex nonlinear relationships and high-dimensional data, which has important advantages for large-scale, high spatiotemporal resolution SLA prediction. When the prerequisite of sufficient SLA data is satisfied, data-driven deep learning models can automatically learn features and perform analyses from the SLA data without the need for manually designing complex physical models (Cui et al., 2023).

Currently, most deep learning methods for SLA prediction tasks focus on network structures based on Recurrent Neural Network (RNN) and its variants (Zhou et al., 2021; Song et al., 2020; Ning et al., 2021; Wang et al., 2022). RNN can remember previous information through their recurrent structure, which helps capture dependencies in time series data (Zaremba et al., 2014; Dudukcu et al., 2023).

* Corresponding author.
E-mail address: jun.bai@ia.ac.cn (J. Bai).

https://doi.org/10.1016/j.isprsjprs.2025.08.005
Received 23 August 2024; Received in revised form 5 June 2025; Accepted 13 August 2025
Available online 8 September 2025

Zhou et al. (2021) proposed a Multilayer Fusion Recurrent Neural Network (MLFrnn) to achieve accurate prediction of SLA. The RNN is employed to capture the dependencies within the SLA time series. Song et al. (2020) proposed a Merged-Long Short Term Memory (Merged LSTM) approach for predicting SLA. This method regards SLA prediction as a time series forecasting problem and divides different regions of the SLA field into small grids for separate predictions. Zhou et al. (2024) developed a Graph-based Memory Recall recurrent neural Network (GMR-Net), achieving spatiotemporal prediction of SLA. The Memory Storage Recall (MSR) module was utilized to capture the medium to long-term temporal dependencies of SLA.

However, RNN and its variants rely on the computation results of the previous step to process the current step. This sequential nature makes it difficult for RNN and its variants to parallelize the calculation, limiting their acceleration on hardware. Besides, RNN is insensitive to spatial features, which also poses challenges to the spatiotemporal prediction of SLA. Therefore, it is necessary to propose a new efficient spatiotemporal network model that can be parallelized for accurate long-term prediction of the SLA.

To address the low computational efficiency of RNN, we employ the Convolutional Neural Network (CNN) as the fundamental computational module to accelerate both training and inference speeds. The SLA data for a specific latitude-longitude region can be regarded as a two-dimensional (2D) time-series image, which is convenient for CNN to process. CNN handles image data through convolution operations, allowing parallel processing at different positions (Li et al., 2021). Convolution operations are very efficient for images or other structured data, and can utilize the parallel computation capabilities of GPUs to enhance computational efficiency (Liu et al., 2022; Li et al., 2022).

Sea eddies are one of the main factors causing SLA. These eddies vary greatly in size and shape (Cui et al., 2022; Jiang et al., 2024). The key to SLA prediction lies in the ability to capture the features of these eddies in the SLA field (Jiang et al., 2023). However, RNN requires flattening the image into a sequence to process 2D data, hindering their ability to effectively capture local spatial features within the eddies. Due to the fixed receptive field of convolutional kernels, typical CNNs also struggle with effective feature extraction when dealing with targets of varying scales.

To solve the problem that ordinary CNNs cannot effectively capture size-varying sea eddies, we adopt a UNet structure as the backbone of the prediction model. UNet employs an encoder–decoder framework. The encoder in UNet utilizes pooling operations to obtain feature maps at different scales, while the decoder uses upsampling and convolution operations to integrate and adjust multi-scale features. This design allows UNet to capture local eddy information in the observed SLA fields across different scales, thereby improving the reasoning ability of future SLA fields. Additionally, the classic encoder–decoder framework of the UNet has been demonstrated to possess excellent feature extraction capabilities and stable generalization performance (Ling et al., 2024; Huo et al., 2024; Liao et al., 2024). Compared to RNNs, the parallel computation of convolutional kernels in the UNet enhances the model's computational efficiency. Compared to Multilayer Perceptrons (MLPs) (Popescu et al., 2009), the weight sharing strategy of convolutional kernels in the UNet significantly reduces the number of parameters.

However, the use of the CNN-based UNet structure also brings two new issues. (1) In the spatial dimension, the local receptive field of CNN limits its ability to process global and long-distance information. (2) In the temporal dimension, CNN may fail to effectively capture time-related dynamic changes in time-series SLA fields.

To address these issues, we specifically design a Temporal-Spatial Transformer Attention (TSTA) module that utilizes attention mechanisms to enhance the ability of the backbone structure to capture global spatial information and temporal variation information. The TSTA module consists of two sub-modules: Spatial Transformer Attention (STA) and Temporal Transformer Attention (TTA). In the spatial dimension, STA utilizes the 2D attention mechanism of Vision Transformer (ViT) (Dosovitskiy et al., 2020) to facilitate global information interaction across the entire SLA space, thereby capturing long-distance dependencies more effectively. In the temporal dimension, TTA creatively combines the time axis with the spatial axis to form a spatiotemporal plane that contains both temporal and partial spatial information. On this spatiotemporal plane, we take advantage of the global information interaction capability of ViT to learn the spatiotemporal dynamic changes of SLA fields. Compared to traditional CNN and RNN, ViT achieves position information encoding and feature interaction through the self-attention mechanism, thus reducing parameter redundancy and computational complexity (Han et al., 2022). Therefore, this treatment can help enhance the efficiency of the proposed model.

At the connection between the encoder and the decoder, the feature maps contain the richest and most critical spatiotemporal information. Based on the TSTA module, a Center TSTA (CTSTA) module is further designed to perform finer-grained attention weighting.

Ultimately, each level of the UNet backbone is equipped with a lightweight TSTA module, and a separate fine-grained CTSTA module is placed at the center position to form the UNet-TSTA model for the SLA prediction task. To the best of our knowledge, this is the first attempt to use the UNet and ViT method for predicting SLA fields.

Datasets provided by the Copernicus Marine Environment Monitoring Service (CMEMS) are adopted to verify the superior performance of our SLA prediction method. These datasets are constructed from three representative regions: South China Sea (SCS), Tropical Western Pacific (TWP), and Asia Pacific Sea (APS). The model is trained in the APS region and independently validated in the SCS, TWP, and APS regions. The results indicate that in all regions, our UNet-TSTA model can achieve an average prediction error of millimeter (mm) level for 7-day prediction, outperforming other advanced methods.

In summary, the main contributions of this paper are highlighted as follows:

- A UNet-TSTA model is employed to efficiently and accurately predict SLA fields for the upcoming days. The multi-level structure of UNet, taken as the main backbone of the model, aids the prediction model in capturing multi-scale sea eddy features. The TSTA utilizes Point-by-Point Vision Transformer (P-ViT) to enhance UNet's ability to extract key spatiotemporal features. The parallel computing capabilities of CNN and ViT facilitate efficient parallel processing on GPUs, thereby improving computational efficiency.
- A TSTA module is proposed, including STA and TTA submodules, to simultaneously enhance the spatiotemporal modeling ability of the prediction model. The STA module utilizes P-ViT to solve the problem of insufficient analytical ability of UNet for long-distance spatial positional dependencies. The TTA module utilizes the freely combined spatiotemporal planes to assist the UNet structure in effectively capturing time-dependent dynamic changes in SLA fields.
- A CTSTA module is proposed to further enhance the spatiotemporal attention fitting ability of TSTA. At the connection between the encoder and decoder, the spatiotemporal information of the feature map is abundant and crucial. The CTSTA module can provide a more fine-grained attention mask here, adaptively allocating more attention weights to the model, thereby achieving flexible attention to different spatiotemporal positions.
- For the task of predicting 7-day SLA fields, the UNet-TSTA model achieves an average Root Mean Squared Error (RMSE) of 0.815 centimeters (cm), 0.714 cm, and 1.140 cm on the SCS, TWP, and APS datasets, respectively. Compared with MLFrnn, one of the most advanced methods in the SLA field, our model's RMSE performance can be improved by 17.3%. When compared to mainstream ConvLSTM methods in recent years, our model exhibits smaller prediction errors and improves computational efficiency by 6.3 times.

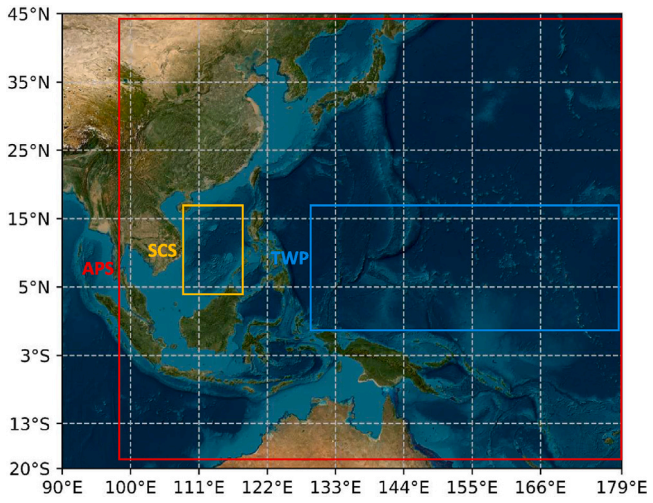

Fig. 1. The overview of SCS, TWP, and APS locations.

2. Data and study region

2.1. Dataset

The dataset used in our study is sourced from the CMEMS https://data.marine.copernicus.eu/product/SEALEVEL_GLO_PHY_L4_MY_008_047/description. Data from satellite altimeters produced SLA daily global grid estimates by Archiving, Validating, and Interpreting Satellite Oceanographic (AVISO) data. The dataset spans from January 1, 1993, to December 31, 2021, encompassing 10,592 days with daily observations. The spatial resolution is 1/4° latitude × 1/4° longitude. We use the SLA fields from January 1, 1993, to December 31, 2015, as the training set, and from January 1, 2016, to December 31, 2021, as the test set. The training set accounts for 80% and the test set accounts for 20% of the dataset. Each independent sample consists of continuous $T_{obs}+T_{pred}$ days of SLA fields, with the first T_{obs} days used as the model input and the subsequent T_{pred} days as the labels. The minimum time interval between each independent sample is 3 days. Both T_{obs} and T_{pred} are adjustable, and in this study, we set T_{obs} = 28 and T_{pred} = 7 by default. For the convenience of neural network computation and evaluation, the values of land areas in the SLA field will be uniformly set to zero.

2.2. Research sea areas

Previous studies often focused on localized coastal areas (Peng and Deng, 2020a,b; Peng et al., 2021, 2024); however, our research extends into the deeper marine interior, analyzing the dynamic variations in SLA on a larger scale. Therefore, three regions in the Pacific Ocean are selected to analyze and validate the performance of our SLA prediction algorithm, namely SCS, TWP, and APS, as shown in Fig. 1.

The SCS region is a semi-enclosed basin connecting the Pacific and Indian Oceans. Due to influences such as monsoons, tides, and topography, this area frequently experiences mesoscale eddies and storm surges, significantly affecting SLA fields (Metzger and Hurlburt, 2001; Nan et al., 2011; Zhao et al., 2014; Zheng et al., 2014). Thus, the SLA fluctuation characteristics in SCS make the validation of our network model more representative. SCS covers the spatial range of 4.875°N–19.625°N, 109.875°E–119.625°E.

The TWP is one of the regions with the most complex circulatory system globally, significantly impacting the global ocean and climate systems (Hu et al., 2021; Whan et al., 2014; Li and Zhou, 2012). Rossby waves in this region can significantly influence sea level heights (Meyers, 1979; Chelton et al., 2003; Kessler, 1990), subsequently affecting global climate change and ocean phenomena movements (Carton et al., 2005; Norris et al., 2013). Therefore, studying SLA fluctuations in the TWP region can better validate the robustness of the proposed prediction model. TWP covers the spatial range of 0.125°N–19.875°N, 130.125°E–179.875°E.

The APS region is one of the main areas where El Niño and La Niña phenomena occur, which significantly impact global ocean and climate system patterns (Yoshida et al., 2007; You et al., 2021; Gao et al., 2019). Studying SLA in this region can improve the prediction accuracy of extreme natural phenomena, helping to mitigate and respond to their potential impacts. We select a subregion covering 19.875°S to 44.875°N and 100.125°E to 179.875°E for our study. Researching the APS region allows for the full utilization of the extensive spatial information of SLA fields beyond the SCS and TWP regions.

Different sea areas have their own respective marine environmental characteristics. By independently testing in these different sea areas, the adaptability of the model can be evaluated under various complex marine environments. During the training process, the parameters learning of the neural network is conducted only for the relatively large-scale APS region. During the testing process, however, evaluations on the prediction performance are conducted independently in each of the three sea areas. This enables us to avoid the high cost of training three different models separately.

3. Methodology

3.1. Network structure

The overall architecture of the UNet-TSTA model is illustrated in Fig. 2. The backbone of the UNet-TSTA model adopts a UNet structure. This UNet backbone primarily consists of an encoder, a decoder, skip connections, and an output block. Both the encoder and decoder incorporate multiple convolutional blocks to construct multi-scale feature representations. To address the UNet backbone's insensitivity to spatiotemporal dynamic features, the TSTA module, equipped with spatiotemporal attention mechanisms, is embedded within each convolutional block.

The encoder comprises five serially connected convolutional modules, each followed by a TSTA module. Following each TSTA module, a max-pooling operation is performed to compress the spatial dimensions of the feature maps. Each convolutional module contains four 2D convolutional layers with a kernel size of 3 × 3, and the number of convolutional kernels increases with the depth of the network.

The junction between the encoder and the decoder is composed of a CTSTA module, which is responsible for applying a more fine-grained attention mechanism and weighting to the encoded feature maps.

The decoder includes five convolutional modules and four TSTA modules. Each convolutional module also contains four 2D convolutional layers with a kernel size of 3 × 3, and the number of convolutional kernels decreases with the depth of the decoder network. Following each convolutional module, an upsampling operation is performed to restore the spatial dimensions of the feature maps.

The encoder and the decoder are connected through skip connections. The feature maps of the encoder and the decoder are aligned in size through padding in the spatial dimension, and integrated together through concatenation in the channel dimension. During the process of gradually restoring the feature maps to the original input size, the decoder can dynamically adjust the attention weights based on the correlation between features in the encoder and the decoder.

At the end of the decoder, there is a convolutional output block that converts the decoded feature maps into predicted SLA fields with a time sequence of a fixed length. This output block has a convolutional kernel size of 1 × 1, and the number of kernels corresponds to the predicted time length.

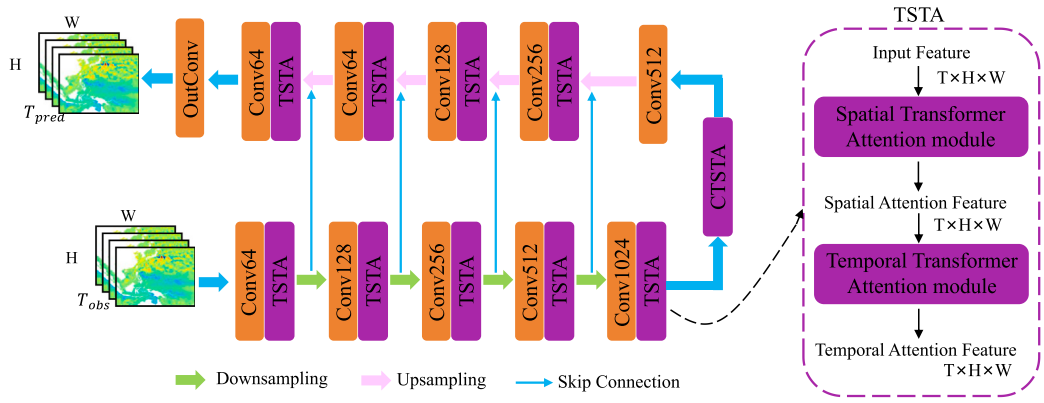

Fig. 2. The overview of the UNet-TSTA method. The T_{obs} represents the time length of observed continuous SLA fields. H and W represent the longitude and latitude ranges of the target area, respectively. The T_{pred} represents the time length of predicted continuous SLA fields. The $ConvN$ represents the convolutional layer with N kernels. TSTA and CTSTA represent the attention module. The $OutConv$ represents the convolutional layer with a kernel size of 1×1, which is used to adjust the prediction time length of output SLA fields.

3.2. Unet backbone structure

Conventional SLA prediction architectures based on RNN and its variants mainly focus on temporal sequence modeling but are weak in capturing spatial dependencies. As an alternative the encoder–decoder architecture of UNet (Ronneberger et al., 2015) can effectively capture hierarchical spatial features from local to global scales. Additionally, the architecture of UNet can help enhance the capability of parallel computations.

Therefore, in this paper, the UNet architecture is adopted as the backbone network. The backbone mainly consists of an encoder, a decoder, skip connections, and an output block. The encoder is responsible for spatiotemporal feature extraction and analysis of observed SLA fields, while the decoder is tasked with accurately predicting future changes. The skip connections help mitigate common issues such as gradient vanishing and information loss during feature transmission. The output block transforms the feature channels into a sequence of SLA fields with a fixed time order.

Since SLA data contains rich temporal sequence information, we choose to concatenate SLA of different times along the channel dimension, forming a 3D cube with dimensions $T \times H \times W$, where T represents the temporal dimension, W stands for the width dimension, and H is the height dimension. The UNet backbone network adopts 2D convolutions in the $H \times W$ dimension to complete feature extraction and reconstruction, while the T dimension is treated as the channel dimension of the feature map.

3.3. TSTA module

Traditional RNN suffers from the issues of low computational efficiency, gradient vanishing or exploding. In contrast, Transformer utilizes self-attention mechanisms, which are good at capturing long-range dependencies, making them more effective for handling long-sequence data (Yang et al., 2024). The ViT utilizes the patch partitioning mechanism to extend the Transformer's ability of analyzing long-range dependencies to the 2D image (Khan et al., 2022). Additionally, ViT can compute information from different spatial positions in parallel, significantly improving efficiency during training and inference stages (Li et al., 2023).

Therefore, this paper constructs a TSTA module based on the ViT, applying attention weighting to the feature maps extracted by the UNet backbone network in both spatial and temporal dimensions. The detailed structure of the TSTA module is shown in Fig. 3. The TSTA module consists of two parts: STA and TTA, which respectively capture the dependencies in the spatial and temporal dimensions of the features.

3.3.1. STA module

As illustrated in Fig. 3(a), the STA employs a P-ViT to capture dependencies among different regions in the spatial dimension of the input feature map. We segment the input feature map into a series of patches and then use a linear layer to convert each patch into a feature vector. These vectors are subsequently fed into the Transformer Encoder.

The process of dividing the feature map into multiple patches is shown in the following equations.

$$p_{m,n}^{hw} = \{X(h,w) | h \in [m \cdot p_h, (m+1) \cdot p_h), w \in [n \cdot p_w, (n+1) \cdot p_w)\}, \quad (1)$$

where X is the input feature map, $p_{m,n}^{hw}$ stands for the segmented patch in spatial dimension, $m = 1, 2, \ldots, M$, $n = 1, 2, \ldots, N$, $p_{m,n}^{hw} \in \mathbb{R}^{T \times p_h \times p_w}$, $X \in \mathbb{R}^{T \times H \times W}$, H and W represent the height and width of the X, respectively. p_h and p_w stand for the height and width of a patch, respectively. $M = \lfloor H/p_h \rfloor$, $N = \lfloor W/p_w \rfloor$, $\lfloor \cdot \rfloor$ represents the floor function, and $i, j, p_h, p_w \in \mathbb{Z}$.

Subsequently, each patch is embedded into a feature vector through the linear layer.

$$\begin{aligned} p_{m,n}^{wh} &= Flatten(p_{m,n}^{hw}), \ p_{m,n}^{hw} \in \mathbb{R}^{1 \times T \cdot p_h \cdot p_w}, \\ X_{m,n}^{hw} &= p_{m,n}^{hw} W_{m,n}^{hw} + b_{m,n}^{hw}, \ X_{m,n}^{hw} \in \mathbb{R}^{1 \times L_{token}}, \end{aligned} \quad (2)$$

where $Flatten(\cdot)$ represents the operation of flattening the multidimensional data, $X_{m,n}^{hw}$ is the embedding result, and $W_{m,n}^{hw}$ and $b_{m,n}^{hw}$ stand for the updatable weight and bias of the linear layer.

The feature vectors from various patches carry effective information of input features at different spatial positions. These feature vectors will be input into the Transformer Encoder for global attention feature extraction. Fig. 3(c) describes the detailed structure of the Transformer Encoder, and the most crucial part is the Multi-Head Attention module.

The Multi-Head Attention module is defined as follows. First, linear transformations are applied to generate the Query (Q), Key (K), and Value (V) for multiple heads.

$$\begin{aligned} Q^i &= X^{hw} W^{Q_i}, \\ K^i &= X^{hw} W^{K_i}, \\ V^i &= X^{hw} W^{V_i}, \end{aligned} \quad (3)$$

where X^{hw} represents the embedding result of patches in spatial dimension. W^{Q_i}, W^{K_i} and W^{V_i} are learnable weight matrices for generating Q^i, K^i and V^i. i stands for the ith head.

Then, the attention feature $head_i$ is calculated by the dot product method. The multi-head attention mechanism allows the model to focus on input features from multiple different perspectives (heads).

$$\begin{aligned} head^i &= Attention(Q^i, K^i, V^i) \\ &= softmax\left(\frac{Q^i (K^i)^T}{\sqrt{d_K}}\right) V^i, \end{aligned} \quad (4)$$

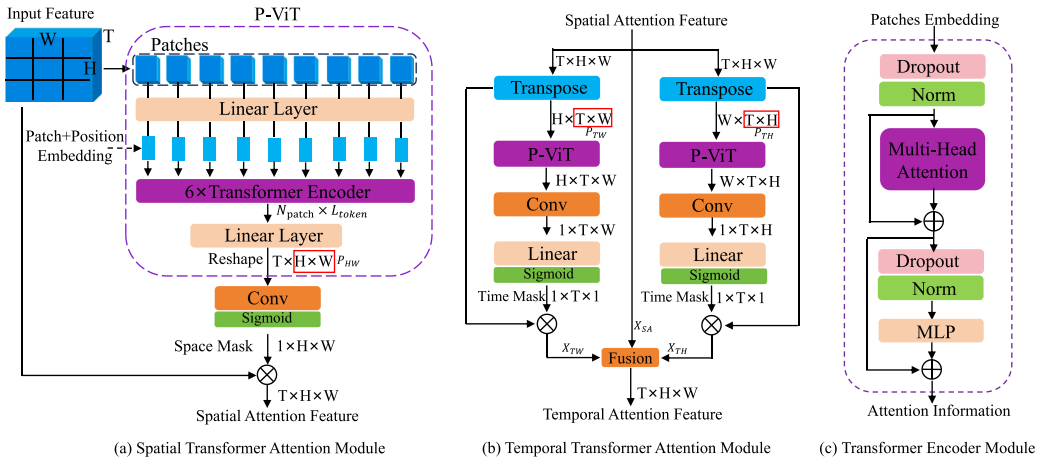

Fig. 3. The detailed structure of the TSTA module. The STA module extracts spatial attention features in the spatial dimension, while the TTA module obtains temporal attention features in the spatiotemporal dimension. The *Norm* represents the layer normalization.

where d_K is the dimension size of K^i, $\sqrt{d_K}$ is used for numerical scaling to avoid gradient vanishing due to excessive inner product values. The softmax function is used to standardize scores so that they can be represented as probability distributions.

Finally, the outputs of all the heads are concatenated and fused by linear operations.

$$\begin{aligned} X^{MH} &= MultiHead(Q,K,V) \\ &= Cat(head^1, head^2, \ldots, head^h)W^O, \end{aligned} \tag{5}$$

where X^{MH} is the Multi-Head Attention results, h stands for the number of heads, and W^O represents a learnable weight matrix that performs a linear transformation on the concatenated multi-head attention output. In our model, h is set to 4. This method allows the model to notice sequences of different parts in different headers, enriching its feature representation capabilities.

In Fig. 3(a), we employ P-ViT to keep the invariance of data dimensions. Unlike the conventional ViT, our P-ViT restores the tokens generated by the Transformer Encoder to the same size as the input feature map through the linear layer and the reshape operation. This measure can establish point-by-point mapping between the output features of the Transformer and the original input features, as shown in the following equations.

$$\begin{aligned} X^{te}_{m,n} &= X^{TE}_{m,n}W^{TE}_{m,n} + b^{TE}_{m,n},\ X^{te}_{m,n} \in \mathbb{R}^{1\times T\cdot p_h \cdot p_w}, \\ X^{p\text{-}vit}_{m,n} &= Reshape(X^{te}_{m,n}),\ X^{p\text{-}vit}_{m,n} \in \mathbb{R}^{T\times p_h \times p_w}, \end{aligned} \tag{6}$$

$$X^{p\text{-}vit}_{hw} = \begin{bmatrix} X^{p\text{-}vit}_{11} & X^{p\text{-}vit}_{12} & \ldots & X^{p\text{-}vit}_{1N} \\ X^{p\text{-}vit}_{21} & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ X^{p\text{-}vit}_{M1} & X^{p\text{-}vit}_{M2} & \ldots & X^{p\text{-}vit}_{MN} \end{bmatrix}, \tag{7}$$

where $X^{p\text{-}vit}_{hw}$ records the result of P-ViT, and $X^{p\text{-}vit}_{hw} \in \mathbb{R}^{T\times H\times W}$, $X^{TE}_{m,n}$ represents the calculation result of the Transformer Encoder for a patch, and $X^{te}_{m,n}$ stands for the result of the linear layer.

Subsequently, a convolution operation with a single convolution kernel is used to compress the temporal dimension of $X^{p\text{-}vit}_{hw}$ to 1, followed by a Sigmoid operation to obtain an attention weighted mask on the spatial dimension. This attention mask multiplies the original input feature map in the spatial dimension to transfer the learned spatial dependencies to the UNet backbone.

3.3.2. TTA module

The Spatial Attention Feature obtained by STA is transmitted to the TTA module to further capture the temporal dependencies of local spatial positions at different time points. As shown in Fig. 3(b), the TTA module consists of two network branches: the P_{HW} branch and the P_{TW} branch. Through the unrestricted combinations of the dimensions T, H, and W, we obtain three different planes: P_{HW}, P_{TW}, and P_{TH}, representing $H\times W$, $T\times W$, and $T\times H$ dimensions, respectively. Among them, the P_{HW} contains only spatial information, while P_{TW} and P_{TH} contain rich spatiotemporal information, exhibiting strong correlations with temporal changes over time.

It is worth noting that typical ViT operations primarily partition patches in the spatial plane (P_{HW}) to capture the relationships between patches in different spatial positions. However, our TTA method innovatively partitions patches in the P_{TW} and P_{TH} planes associated with time, and acquires a dependency matrix of local spatial and temporal relations through P-ViT. Therefore, TTA includes two similar network branches that calculate the feature distribution relationships in the P_{TW} and P_{TH} planes, respectively.

The split patch operation in the P_{TW} and P_{TH} spatiotemporal planes is shown below:

$$p^{tw}_{m,n} = \{X(t,w) | t \in [m\cdot p_t, (m+1)\cdot p_t), w \in [n\cdot p_w, (n+1)\cdot p_w)\}, \tag{8}$$

$$p^{th}_{m,n} = \{X(t,h) | t \in [m\cdot p_t, (m+1)\cdot p_t), h \in [n\cdot p_h, (n+1)\cdot p_h)\}, \tag{9}$$

where $p^{tw}_{m,n}$ and $p^{th}_{m,n}$ represent two patches in the P_{TW} and P_{TH} spatiotemporal planes, respectively. $p^{tw}_{m,n} \in \mathbb{R}^{H\times p_t \times p_w}$, and $p^{th}_{m,n} \in \mathbb{R}^{W\times p_t \times p_h}$.

In the original ViT, the width and height of the input image are the same, so patches are typically square. However, in the UNet encoder, the T dimension of the feature map gradually increases while the W and H dimensions gradually decrease; in the decoder of UNet, the opposite is true. Therefore, the length of the T dimension often significantly differs from the lengths of the W and H dimensions, making it unsuitable for square patch partitioning. To address this issue, in our P-ViT method, we separately configure the height and width of the patches for each layer of the UNet, making them appropriately sized rectangles. Table 1 shows the detailed configuration of the patch size. When the channel dimension T of the feature map increases, the slice length p_t of the patch will increase appropriately, and vice versa. The slice length p_t must be divisible by T. The setting rules for other dimensions are the same.

In the P_{TW} branch of the TTA module, the H dimension of the output from P-ViT is compressed to 1 through a convolution layer, and the W dimension is compressed to 1 through a linear layer. Then, the weight mask information containing only the T dimension is obtained through the Sigmoid function. Finally, this weight mask information is applied to the input feature map of the TTA module through multiplication. The operations in the P_{TH} branch are similar. The attention information obtained from the P_{TW} branch and the P_{TH} branch is then integrated into the backbone network's feature map through the Fusion module.

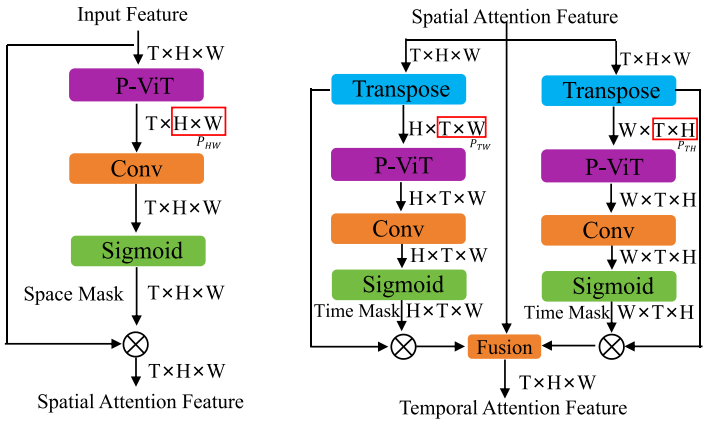

Fig. 4. The detailed structure of the CTSTA module.

Table 1
Patch size configuration for each layer of the UNet-TSTA. Feature Size = (T, H, W). Patch Size = (p_t, p_h, p_w), representing the slice length of each patch in the T, H, and W dimensions, respectively.

Module	Layer	Feature size	Patch size
In	0	(28,320,260)	–
Encoder	1	(64,320,260)	(4,32,26)
	2	(128,160,130)	(8,8,10)
	3	(256,80,65)	(16,8,5)
	4	(512,40,32)	(32,4,2)
	5	(1024,20,16)	(64,2,2)
CTSTA	6	(1024,20,16)	(64,2,2)
Decoder	7	(512,20,16)	(32,2,2)
	8	(256,40,32)	(16,4,2)
	9	(128,80,65)	(8,8,5)
	10	(64,160,130)	(4,8,10)
Out	11	(7,320,260)	–

The definition of the Fusion module is shown in Eq. (10).

$$\begin{aligned} X_F &= Fusion(X_{TW}, X_{TH}, X_{SA}) \\ &= Conv_{1\times1}(Cat(X_{TW} + X_{TH}, X_{SA})) \\ &= Conv_{1\times1}(X_{cat}), \end{aligned} \tag{10}$$

where X_{TW} and X_{TH} stand for the attention feature obtained from the P_{TW} branch and the P_{TH} branch of the TTA module, respectively, X_{SA} is the spatial attention feature, X_{TW}, X_{TH} and $X_{SA} \in \mathbb{R}^{T\times H\times W}$, $Cat(\cdot)$ represents the concatenation operation in the channel dimension, X_{cat} stands for the concatenation result of the $Cat(\cdot)$, $X_{cat} \in \mathbb{R}^{2T\times H\times W}$, $Conv_{1\times1}(\cdot)$ is a convolutional layer with a kernel size of 1×1, X_F represents the result of feature fusion, which is the output of TTA module, and $X_F \in \mathbb{R}^{T\times H\times W}$.

3.4. CTSTA module

The junction of the UNet encoder and the decoder is the central module of the entire SLA prediction model. Here, the spatial information of the input feature maps is compressed to the minimum, while the temporal dimension is expanded to its maximum. Consequently, these feature maps contain refined spatial information and abundant temporal sequence variation information. Applying a more fine-grained attention mechanism weighting to these feature maps can significantly enhance SLA prediction accuracy.

The CTSTA module is proposed based on TSTA to perform a more fine-grained attention mechanism weighting on the feature maps at the central part of the UNet backbone network. The detailed structure of the CTSTA module is illustrated in Fig. 4. The CTSTA module consists of two parts: Center STA (CSTA) and the Center TTA (CTTA), which respectively capture the spatial and temporal dependencies of the input features. However, in the CSTA module, the number of convolution kernels that generate the spatial mask is adjusted to equal the number of T dimensions. That is, the spatial mask of CSTA has a dimension of $T \times W \times H$, instead of $1 \times W \times H$ in STA. Compared to the STA module, the weight distribution in the spatial mask here can more finely map to each data point of the input features, enabling more accurate localization and identification of key structures in the feature maps. Similarly, in the CTTA module, the linear layers used to compress the W or H dimensions are removed. As a result, the temporal mask of CTTA is more fine-grained, having a dimension of $H \times T \times W$ or $W \times T \times H$, instead of $1 \times T \times 1$ in TTA.

Since the CTSTA module requires a higher data density for the attention mask compared to TSTA, the length of the hidden token vectors in the transformer encoder should be appropriately increased to accommodate more complex and flexible weight variations. In this paper, the token vector length (L_{token}) of CTSTA is set to 768, while that of TSTA is set to 128. Due to the substantial number of parameters in the CTSTA module, we utilize a single CTSTA module only in the central area of the UNet backbone structure to enhance the prediction precision. In contrast, in the UNet encoder and decoder, the TSTA modules with fewer parameters are employed to dynamically adjust the focused features through a layer-by-layer attention mechanism.

3.5. Loss function and evaluation indicators

The Mean Squared Error (MSE) is a convex function, which makes it easier to find the global optimum during the optimization process, avoiding local optima. Therefore, we choose MSE as the loss function for model training, as shown in Eq. (11).

$$loss_{mse} = \frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2, \tag{11}$$

where $\hat{y}_i$ represents the predicted result of the model for the ith sample, y_i stands for the true value of the ith sample, $\|\cdot\|_2$ represents $L2$ norm, n is the number of samples, and $loss_{mse}$ represents the MSE loss value.

The Adaptive Moment Estimation (Adam) optimizer dynamically adjusts the learning rate based on the estimates of the first and second moments of the parameters' gradients, allowing different parameters to have different learning rates, which helps to accelerate the learning process and improve the model's convergence speed. Thus, we adopt the Adam optimizer to update the model parameters.

Additionally, we use other commonly employed evaluation metrics in SLA prediction, RMSE and Mean Absolute Error (MAE), to assess the

Table 2
Comparison results with other advanced prediction methods in SCS area.

Prediction (day)		1st	2nd	3rd	4th	5th	6th	7th	Average
FCN (Long et al., 2015)	MAE (cm)	2.023	2.029	2.062	2.128	2.212	2.318	2.435	2.172
Motion RNN (Wu et al., 2021)		0.828	1.062	1.317	1.580	1.844	2.101	2.349	1.583
ConvLSTM (Su et al., 2020)		**0.178**	0.322	0.496	0.702	0.932	1.177	1.429	0.748
MLFrnn (Zhou et al., 2021)		0.236	0.374	0.533	0.713	0.909	1.114	1.325	0.743
SmaAtUNet (Trebing et al., 2021)		0.255	0.337	0.476	0.650	0.841	1.042	1.242	0.692
Ours		0.194	**0.294**	**0.427**	**0.582**	**0.752**	**0.933**	**1.120**	**0.614**
FCN (Long et al., 2015)	RMSE (cm)	2.745	2.759	2.795	2.870	2.969	3.101	3.251	2.927
Motion RNN (Wu et al., 2021)		1.326	1.576	1.864	2.181	2.509	2.837	3.158	2.207
Merged LSTM (Song et al., 2020)		0.280	0.590	0.880	1.200	1.600	–	–	–
ConvLSTM (Su et al., 2020)		**0.238**	0.425	0.654	0.927	1.232	1.558	1.893	0.990
MLFrnn (Zhou et al., 2021)		0.324	0.506	0.714	0.946	1.200	1.466	1.739	0.985
SmaAtUNet (Trebing et al., 2021)		0.339	0.447	0.630	0.861	1.114	1.378	1.644	0.916
Ours		0.259	**0.390**	**0.565**	**0.770**	**0.997**	**1.238**	**1.487**	**0.815**

model's predictive performance, as shown in Eqs. (12) and (13).

$$RMSE = \frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2, \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_1, \tag{13}$$

where y_i represents the ith true value, $\hat{y}_i$ stands for the ith predicted value, $\|\cdot\|_1$ denotes $L1$ norm, and n represents the number of samples.

4. Experimental results

In this section, we conduct detailed comparative experiments to verify the superiority of our proposed method. For the convenience of readers, the data unit in this paper is set to cm by default.

4.1. Comparative experiments with other advanced methods

To validate the superiority of the proposed UNet-TSTA method, we conduct comparative experiments with other advanced prediction methods, including Merged LSTM (Song et al., 2020), MLFrnn (Zhou et al., 2021), ConvLSTM (Su et al., 2020), SmaAtUNet (Trebing et al., 2021), Motion RNN (Wu et al., 2021), and FCN (Long et al., 2015). The prediction duration for models is set to 7 days. The longer time comparison results can refer to Fig. 9 and Section 4.5. The experimental results for MLFrnn and Merged LSTM methods in the SCS region are obtained directly from the original papers (Zhou et al., 2021; Song et al., 2020), as their data sources and the cropped area are the same as ours. For the lack of publicly available source code, the results of the other two areas (TWP and APS) are absent for these two methods. The original paper of the Merged LSTM method only conducted SLA prediction for the next 5 days, coinciding with the results presented in Table 2.

Tables 2–4 present a comparison of different prediction models. Our UNet-TSTA model achieves the lowest RMSE and MAE across three distinct sea areas. Among the compared models, the average prediction errors of the ConvLSTM, the MLFrnn, and the SmaAtUNet are all within the mm range. Compared to one of the state-of-the-art methods, MLFrnn, which is specifically designed for the SLA prediction task, our model shows a 17.3% improvement in prediction performance on average. The SmaAtUNet method employs a model architecture similar to our UNet-TSTA approach, utilizing a combination of UNet and an Attention module. Therefore, its prediction performance is the closest to ours. This similarity suggests that the integration of attention mechanisms within a UNet framework can lead to notable improvements in prediction precision for SLA tasks.

We attempt to employ the Fully Convolutional Network (FCN) approach (Long et al., 2015) for the SLA prediction task. However, the FCN exhibits significant prediction errors. Compared to the UNet structure, FCN lacks the skip connection architecture. The downsampling and upsampling operations in the encoder and the decoder may result in the loss of critical spatial information, leading to less refined prediction outcomes. The absence of skip connections prevents effective compensation for these spatial information losses. Furthermore, the lack of skip connections prevents the effective integration of lower-level and higher-level features. Additionally, FCN lacks transformer attention mechanisms, resulting in its inability to effectively capture dynamically spatiotemporal information. Therefore, relying solely on CNN structures proves insufficient for accurately predicting the SLA field.

The Motion RNN (Wu et al., 2021) method is originally designed to predict spatiotemporal variations in visual videos by simultaneously capturing transient changes and motion trends. The reliance on RNN-based frameworks inherently limits their ability to effectively extract local spatial features, which are crucial in tasks like SLA prediction. Furthermore, the method's design for natural image sequences might not seamlessly transfer to SLA field prediction, where the dynamics and characteristics of the data differ significantly. Thus, despite its capabilities in capturing transient changes and motion trends in natural visual videos, the Motion RNN method may not be optimally suited for the specific demands of SLA prediction tasks.

The Merged LSTM (Song et al., 2020) method employs a three-layer LSTM approach to accomplish the SLA prediction task. This method treats SLA prediction as a time-series forecasting problem and divides the SLA field into several small subgrids for independent prediction. This grid division approach fragments the originally complete 2D spatial information, making it difficult for the network model to capture the global dependencies between subgrids. LSTM is primarily designed for handling time-series data, and when applied to the SLA field processing, it struggles to effectively capture and retain the spatial features of SLA fields.

The ConvLSTM (Su et al., 2020) integrates the strengths of both CNN and LSTM, effectively capturing spatial and temporal features when processing spatiotemporal data. Consequently, its predictive performance surpasses that of both FCN and RNN methods. The ConvLSTM adopts a recurrent prediction strategy, predicting the SLA field for only 1 day at a time and using the predicted result as input to sequentially predict the following days. When predicting the first day, ConvLSTM focuses on the current time step and does not consider next multi-step prediction issues, which leads to slightly better short-term prediction results compared to our UNet-TSTA method. However, as the number of prediction days increases, the errors are gradually accumulated, leading to a significant decline in the next days' performance. This is due to the fact that the trend of SLA changes is difficult to manifest in the short term, and ConvLSTM tends to generate results similar to the last day in the input at each step, which impacts the overall accuracy of the prediction.

In contrast, our proposed method uses a one-step prediction strategy for the next 7 days, avoiding the error propagation issues that may arise in step-by-step predictions. At the same time, the UNet-TSTA introduces a spatiotemporal attention mechanism based on P-ViT, allowing the

Table 3
Comparison results with other advanced prediction methods in TWP area.

Prediction (day)		1st	2nd	3rd	4th	5th	6th	7th	Average
FCN		1.443	1.452	1.502	1.585	1.684	1.785	1.891	1.620
Motion RNN		0.512	0.766	1.028	1.287	1.536	1.771	1.991	1.270
ConvLSTM	MAE (cm)	**0.152**	0.291	0.460	0.653	0.860	1.071	1.279	0.681
SmaAtUNet		0.223	0.293	0.428	0.590	0.756	0.924	1.092	0.615
Ours		0.161	**0.252**	**0.376**	**0.519**	**0.673**	**0.831**	**0.988**	**0.543**
FCN		2.107	2.116	2.157	2.232	2.326	2.439	2.561	2.277
Motion RNN		0.696	1.017	1.357	1.696	2.024	2.335	2.626	1.679
ConvLSTM	RMSE (cm)	**0.201**	0.382	0.605	0.861	1.134	1.410	1.680	0.896
SmaAtUNet		0.287	0.380	0.561	0.777	0.998	1.220	1.441	0.809
Ours		0.210	**0.330**	**0.493**	**0.683**	**0.887**	**1.095**	**1.302**	**0.714**

Table 4
Comparison results with other advanced prediction methods in APS area.

Prediction (day)		1st	2nd	3rd	4th	5th	6th	7th	Average
FCN		2.396	2.396	2.448	2.528	2.621	2.720	2.820	2.561
Motion RNN		1.049	1.317	1.590	1.863	2.127	2.377	2.614	1.848
ConvLSTM	MAE (cm)	**0.182**	0.346	0.545	0.772	1.014	1.261	1.504	0.803
SmaAtUNet		0.299	0.381	0.549	0.728	0.916	1.126	1.310	0.758
Ours		0.202	**0.311**	**0.460**	**0.631**	**0.812**	**0.994**	**1.175**	**0.655**
FCN		4.117	4.128	4.173	4.255	4.353	4.475	4.605	4.301
Motion RNN		1.899	2.244	2.616	3.007	3.394	3.767	4.121	3.007
ConvLSTM	RMSE (cm)	**0.336**	0.626	0.944	1.321	1.720	2.116	2.495	1.373
SmaAtUNet		0.496	0.645	0.931	1.257	1.583	1.901	2.209	1.289
Ours		0.338	**0.531**	**0.803**	**1.113**	**1.431**	**1.738**	**2.031**	**1.140**

model to better capture long-range temporal dependencies and cross-space correlations inherent in the SLA field. This enables more accurate judgments of SLA changes over multiple future time steps. Experimental results show that as the prediction days progress, the prediction performance of UNet-TSTA surpasses ConvLSTM as early as the second day and demonstrates a significant advantage in predictions for 2-7th days. Additionally, Table 9 demonstrates the computational efficiency and parameter advantages of our UNet-TSTA method, further highlighting its superior performance compared to ConvLSTM.

The MLFrnn (Zhou et al., 2021) designs a multi-layer fusion cell to simultaneously extract and integrate spatiotemporal features based on RNN and convolutions. However, the inherent problem of RNN in effectively utilizing local or global image features remains unresolved. In contrast, our TSTA module utilizes the multi-head attention mechanism of P-ViT and the freely combinable spatiotemporal planes, capturing complex long-range spatiotemporal dependencies. Moreover, P-ViT's Transformer-based modular design allows for flexible adjustment of model size, number of layers, and attention heads, offering greater flexibility.

The SmaAtUNet (Trebing et al., 2021) adopts a structure combining UNet and CBAM to perform spatiotemporal prediction. The CBAM (Woo et al., 2018) module utilizes both spatial and temporal attention mechanisms, which is similar to our TSTA module. The results show that both the SmaAtUNet method and the UNet-TSTA method perform well in prediction tasks. This indicates that attention mechanisms in both time and space are well-suited to solving SLA prediction problems. However, CBAM's spatial and channel attention features are obtained through convolutional and linear layers, whereas our TSTA module acquires attention features via a sophisticated spatiotemporal P-ViT mechanism. The capability of ViT in global feature integration often surpasses that of conventional convolutional or linear layers. On the other hand, SmaAtUNet uses a depthwise separable convolution-based UNet as its backbone. While this reduces the number of parameters and computational load, it can also lead to insufficient expressive capability. Therefore, the prediction results obtained by the UNet-TSTA method are more accurate.

The comparison between the predicted SLA field and the actual SLA field can provide a tangible measure of prediction quality in Figs. 5, 6, and 7. These figures respectively display the predicted SLA fields by the UNet-TSTA model for the SCS, TWP, and APS sea areas. The similarity between the real SLA field and the predicted SLA field is relatively high. This proves that our UNet-TSTA model has successfully captured the dynamic change trends of the SLA field to a certain extent.

In Fig. 5-(c), in the southeast corner of the SCS area, there is a land area. In the coastal regions close to the land, the prediction errors are significantly higher than those in the central regions of the SCS. This is largely due to the reason that, when ocean currents encounter land, they undergo changes such as bifurcation and convergence, forming complex circulation patterns. Additionally, due to the blocking and constraints of the land topography, tidal waves experience reflection, refraction, and interference during their propagation. The superposition of tidal waves and circulation patterns from different directions makes the variation patterns of the SLA field in the nearshore areas difficult to predict accurately.

Additionally, from the comparison between Figs. 5-(c) and 6-(c), it is clear that the prediction errors in the SCS are significantly higher than those in the TWP. The results in Tables 2 and 3 also show the same trend. Under the same condition of using our UNet-TSTA model, the average MAE in SCS area is 13.07% higher than that in the TWP area. This is because the SCS sea area has a complex seabed topography, including basins, trenches, ridges, and continental shelves. In contrast, the topography of the TWP region is relatively more open and gentle, and the movement patterns of the ocean currents are easier to grasp. The numerous islands around the SCS can block and diffract ocean currents and waves, altering the local oceanic dynamic environment. However, some areas of the TWP have fewer islands, and the oceanic dynamic environment is relatively simpler. Therefore, the SLA prediction errors are relatively lower in the TWP region.

4.2. Ablation study

To demonstrate the performance improvement brought by the different modules for the SLA prediction task, we conduct specialized ablation experiments. Tables 5–7 present the experiment results. "UNet" represents the common UNet structure without any auxiliary attention modules. "UNet + TSTA" means that for each convolutional module in the UNet structure, a TSTA module is used to assist the UNet backbone structure in feature extraction and prediction of SLA fields. Additionally, we add a CTSTA module at the connection position between the

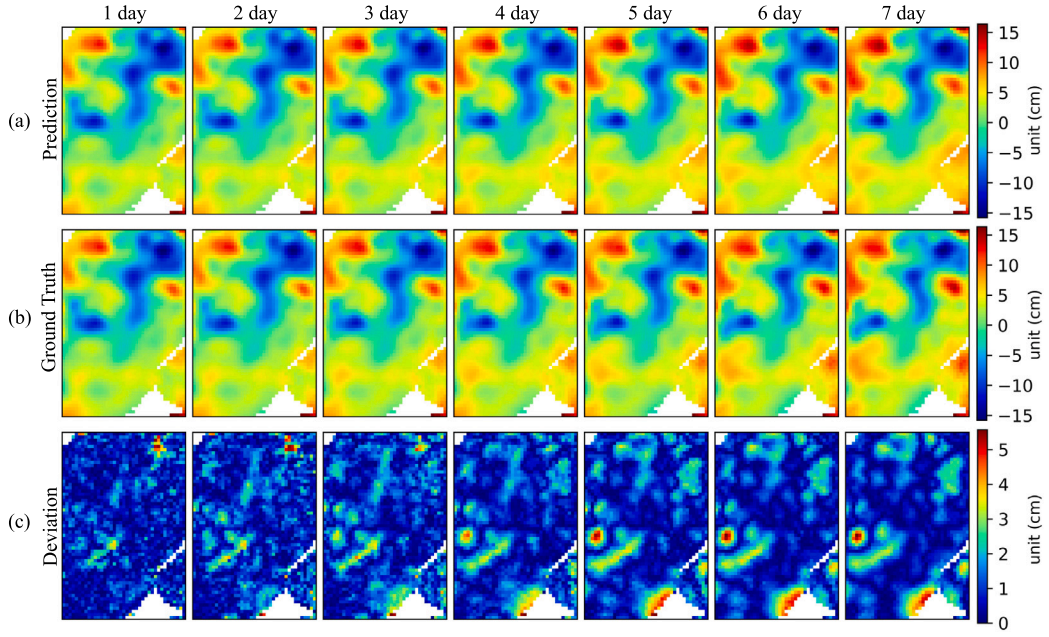

Fig. 5. The prediction results of the UNet-TSTA model for future 7-day SLA fields in SCS area. (a) the predicted SLA field. (b) the real SLA field. (c) the absolute error between the predicted and real SLA field.

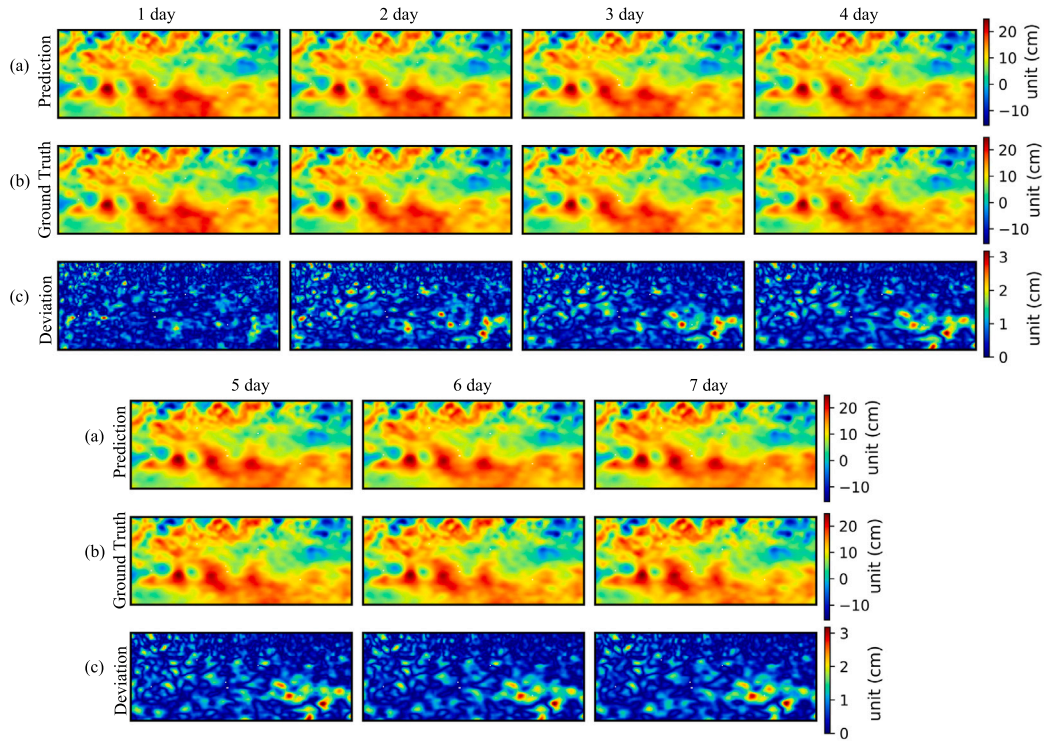

Fig. 6. The prediction results of the UNet-TSTA model for future 7-day SLA fields in TWP area. (a) the predicted SLA field. (b) the real SLA field. (c) the absolute error between the predicted and real SLA field.

Table 5
The TSTA module ablation experiment results in SCS area.

Prediction (day)		1st	2nd	3rd	4th	5th	6th	7th	Average
UNet		0.259	0.351	0.483	0.640	0.814	0.999	1.192	0.676
UNet+TSTA	MAE (cm)	0.202	0.299	0.429	0.584	0.757	0.943	1.136	0.621
UNet+TSTA+CTSTA		**0.194**	**0.294**	**0.427**	**0.582**	**0.752**	**0.933**	**1.120**	**0.614**
UNet		0.376	0.480	0.643	0.844	1.072	1.315	1.569	0.899
UNet+TSTA	RMSE (cm)	0.269	0.396	0.569	0.774	1.004	1.250	1.504	0.823
UNet+TSTA+CTSTA		**0.259**	**0.390**	**0.565**	**0.770**	**0.997**	**1.238**	**1.487**	**0.815**

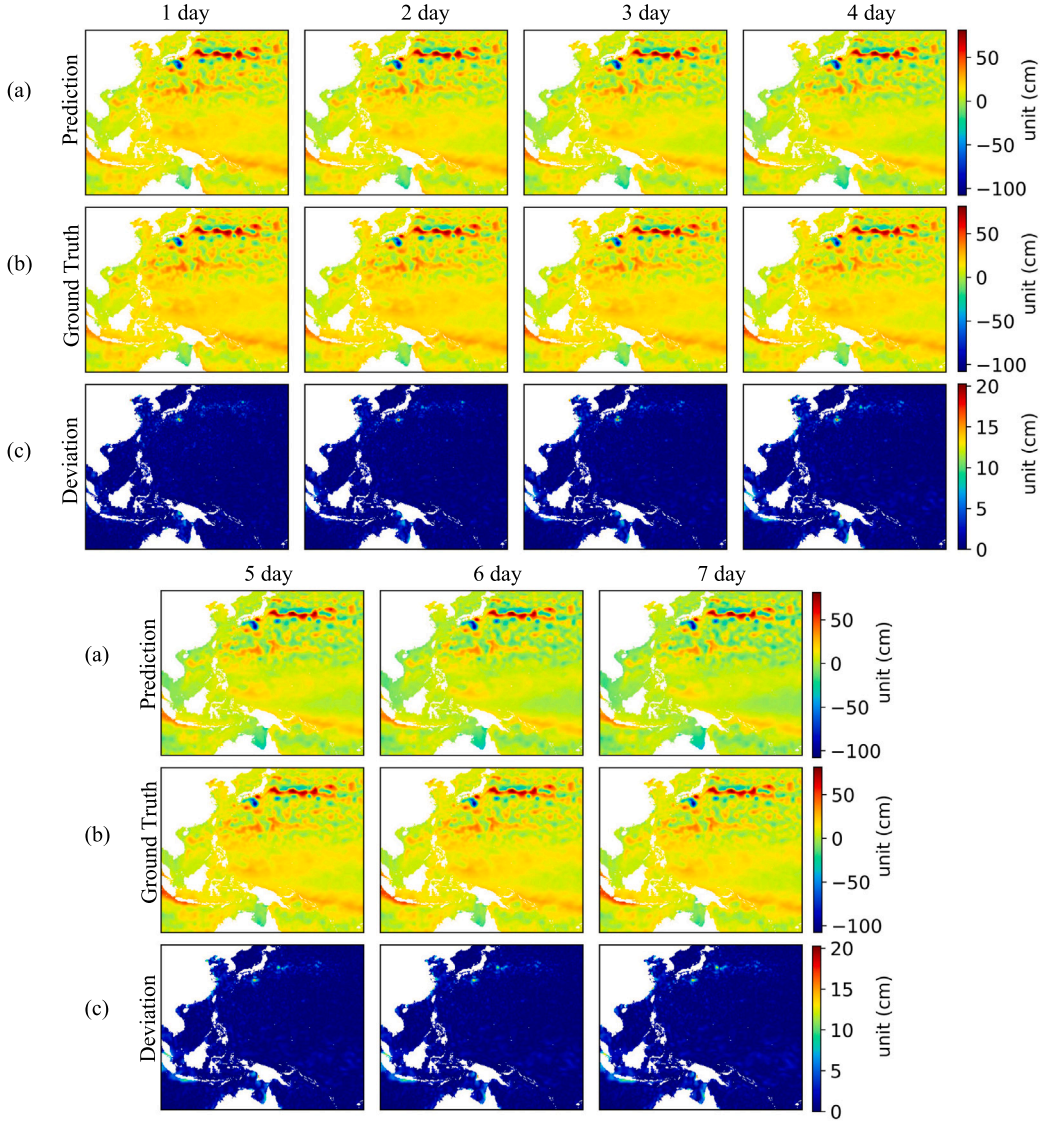

Fig. 7. The prediction results of the UNet-TSTA model for future 7-day SLA fields in APS area. (a) the predicted SLA field. (b) the real SLA field. (c) the absolute error between the predicted and real SLA field.

Table 6
The TSTA module ablation experiment results in TWP area.

Prediction (day)		1st	2nd	3rd	4th	5th	6th	7th	Average
UNet		0.179	0.276	0.413	0.564	0.725	0.884	1.045	0.583
UNet+TSTA	MAE (cm)	0.167	0.256	0.381	0.528	0.685	0.846	1.004	0.552
UNet+TSTA+CTSTA		**0.161**	**0.252**	**0.376**	**0.519**	**0.673**	**0.831**	**0.988**	**0.543**
UNet		0.235	0.360	0.540	0.741	0.953	1.163	1.375	0.766
UNet+TSTA	RMSE (cm)	0.217	0.334	0.500	0.693	0.901	1.112	1.320	0.725
UNet+TSTA+CTSTA		**0.210**	**0.330**	**0.493**	**0.683**	**0.887**	**1.095**	**1.302**	**0.714**

Table 7
The TSTA module ablation experiment results in APS area.

Prediction (day)		1st	2nd	3rd	4th	5th	6th	7th	Average
UNet		0.277	0.374	0.522	0.690	0.871	1.051	1.232	0.717
UNet+TSTA	MAE (cm)	0.211	0.317	0.466	0.639	0.822	1.008	1.190	0.665
UNet+TSTA+CTSTA		**0.202**	**0.311**	**0.460**	**0.631**	**0.812**	**0.994**	**1.175**	**0.655**
UNet		0.471	0.616	0.861	1.152	1.455	1.750	2.037	1.192
UNet+TSTA	RMSE (cm)	0.349	0.540	0.813	1.124	1.444	1.754	2.049	1.153
UNet+TSTA+CTSTA		**0.338**	**0.531**	**0.803**	**1.113**	**1.431**	**1.738**	**2.031**	**1.140**

encoder and the decoder to form the final "UNet + TSTA + CTSTA" structure.

The experimental results in Tables 5–7 demonstrate that the use of the TSTA module improves the model's average prediction performance

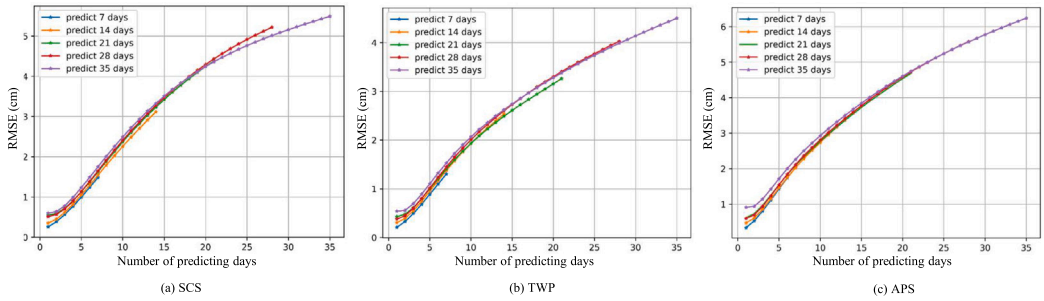

Fig. 8. Experimental results of predicting time length. (a), (b) and (c) represent the experimental results of RMSE in SCS, TWP, and APS areas, respectively.

in the SCS, TWP, and APS sea areas by 8.45%, 5.35%, and 3.38%, respectively. This confirms that the dual attention mechanism in both temporal and spatial dimensions enhances the UNet backbone's ability to analyze and predict SLA fields. Furthermore, after adding the CTSTA module to the prediction model, the average prediction performance improved by 9.34%, 6.79%, and 4.51% in the SCS, TWP, and APS sea areas, respectively, compared to the initial UNet backbone structure. This indicates that incorporating a finer-grained spatiotemporal attention mechanism at the connection position between the encoder and the decoder helps enhance the model's ability to capture and predict the spatiotemporal dynamic changes of the entire SLA fields.

The UNet backbone achieves a 7-day average MAE of less than 1 cm, demonstrating that embedding the temporal dimension into the feature channel dimension of the 2D UNet is highly effective for multi-step prediction of SLA fields. Due to the already low error of the UNet backbone, the reduction magnitude in MAE and RMSE after introducing TSTA modules may appear modest. However, compared to the baseline UNet, the UNet-TSTA model achieves MAE reduction ratios of 9.17%, 6.86%, and 8.67% in the SCS, TWP, and APS regions, respectively. These results indicate that the incorporation of spatiotemporal attention mechanisms leads to a significant improvement. Furthermore, as the prediction time extends, the advantage of the UNet-TSTA model will become increasingly evident due to cumulative error effects. The spatiotemporal attention mechanism in the TSTA module effectively extends the model's valid forecasting period, which is particularly valuable for long-term ocean system analysis.

4.3. Performance over longer predicting time ranges

To explore the predictive performance limits of the UNet-TSTA model, we attempt to predict SLA fields over longer time ranges. We adjust the number of convolution kernels in the output layer of the UNet-TSTA model to control the length of the prediction time (T_{pred}). To control the variables of the comparison experiment, the length of time for input SLA fields is uniformly set to 28 days, while the output lengths of time are set to 7, 14, 21, 28, and 35 days, respectively.

Fig. 8 shows the experimental results for the three different regions. By observing the RMSE curves for different prediction days, we find that the fewer the days predicted at a time, the smaller the prediction error of the model. As the number of prediction days increased, the model's forecasting ability for the initial days gradually deteriorated. This is because when computing the MSE loss, the errors of SLA fields at closer time steps contribute smaller values compared to those farther away. Therefore, errors in SLA fields at distant time steps have a greater impact on the overall MSE loss. To minimize the average MSE loss across all time steps, the optimizer tends to adjust gradients towards directions that reduce the loss at those farther time steps. This may also lead to the optimizer overlooking variations in the loss at closer time steps.

4.4. Performance over different observation time ranges

In this experiment, we investigate the effect of the observation period length T_{obs} on the predictive performance. We only vary the input parameter T_{obs} without altering the model's prediction parameter T_{pred}. Table 8 shows the prediction performance of the model with different observation periods.

When the number of observation days is too small, the model cannot capture the complete variation process of the SLA field from the input data, resulting in poor performance. Conversely, when the number of observation days is too large, the input information becomes redundant, making it difficult for the model to identify key variation intervals within the excessively long time series. From Table 8, we can observe that the prediction model's RMSE and MAE tend to be lower when the observation period is over 28 days. And this consistent outcome is observed across the three different experimental sea areas. Therefore, we select an optimal observation period of 28 days, which yields the best experimental results, as the appropriate input duration for the model.

The length of input time (T_{obs}) for the proposed UNet-TSTA model is determined by the number of input channels in the first convolutional layer. Therefore, the input time window of the model can be flexibly adjusted according to specific application scenarios. In practical applications where SLA data are at risk of high-frequency missing values – such as satellite observation interruptions or transmission failures – a model with a 7-day input window is more likely to meet data completeness requirements. On the other hand, if the historical SLA data is relatively complete, employing a model with a 28-day input window can offer more reliable predictive accuracy.

4.5. Further comparison experiments with other advanced methods over long prediction time ranges

To validate the superior performance of our UNet-TSTA method in long-term SLA prediction, we conduct an experiment to assess its extreme performance in predicting longer durations compared to other advanced prediction methods. Based on the experimental results in Fig. 8, it is evident that predicting long-term SLA fields in a single output adversely affects early-stage performance of the predictions. Therefore, we use the 7-day prediction model and attempt to use the predicted SLA fields together with the observed SLA fields as input to predict 14 more days, taking a rolling strategy to obtain a prediction result of up to 21 days.

Fig. 9 illustrates how the prediction performance of different models varies with increasing prediction days, which is consistent with the experimental results of Tables 2–4 when only predicting 7 days. The RMSE performance curve of our UNet-TSTA method consistently remains below that of other methods, indicating its ability to maintain effective predictions over longer durations. The RMSE performance curves of MLFrnn, ConvLSTM, and SmaAtUNet methods demonstrate good and stable performance. Compared to the state-of-the-art MLFrnn method, the performance advantage of our UNet-TSTA model increases with the number of prediction days.

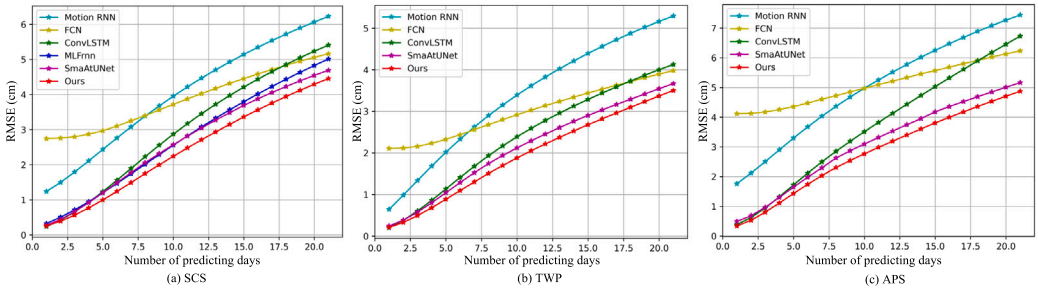

Fig. 9. Comparison results with other advanced methods in different sea areas.

Table 8
Observation days experiment results in different areas.

Prediction (days)			1st	2nd	3rd	4th	5th	6th	7th	Average
SCS	Observation 7 days	MAE (cm)	0.223	0.318	0.451	0.607	0.783	0.970	1.165	0.645
	Observation14 days		0.198	0.302	0.437	0.595	0.769	0.955	1.148	0.629
	Observation 28 days		**0.194**	**0.294**	**0.427**	**0.582**	**0.752**	**0.933**	**1.120**	**0.614**
	Observation 56 days		0.202	0.300	0.435	0.594	0.769	0.956	1.150	0.630
	Observation 112 days		0.206	0.309	0.445	0.606	0.785	0.974	1.170	0.642
	Observation 7 days	RMSE (cm)	0.297	0.418	0.593	0.802	1.036	1.285	1.545	0.854
	Observation14 days		0.265	0.400	0.580	0.789	1.020	1.266	1.523	0.835
	Observation 28 days		**0.259**	**0.390**	**0.565**	**0.770**	**0.997**	**1.238**	**1.487**	**0.815**
	Observation 56 days		0.268	0.398	0.579	0.791	1.025	1.275	1.535	0.839
	Observation 112 days		0.274	0.407	0.588	0.802	1.038	1.290	1.548	0.849
TWP	Observation 7 days	MAE (cm)	0.194	0.278	0.401	0.544	0.698	0.854	1.010	0.568
	Observation14 days		**0.160**	0.255	0.381	0.525	0.680	0.839	0.998	0.548
	Observation 28 days		0.161	**0.252**	**0.376**	**0.519**	**0.673**	**0.831**	**0.988**	**0.543**
	Observation 56 days		0.162	0.256	0.387	0.537	0.697	0.859	1.016	0.559
	Observation 112 days		0.169	0.263	0.392	0.541	0.699	0.861	1.022	0.564
	Observation 7 days	RMSE (cm)	0.253	0.360	0.523	0.714	0.919	1.125	1.331	0.747
	Observation14 days		**0.209**	0.334	0.502	0.693	0.899	1.110	1.319	0.724
	Observation 28 days		0.210	**0.330**	**0.493**	**0.683**	**0.887**	**1.095**	**1.302**	**0.714**
	Observation 56 days		0.212	0.336	0.509	0.709	0.921	1.134	1.343	0.738
	Observation 112 days		0.221	0.345	0.514	0.712	0.921	1.133	1.344	0.741
APS	Observation 7 days	MAE (cm)	0.239	0.333	0.477	0.645	0.824	1.004	1.183	0.672
	Observation14 days		0.206	0.315	0.466	0.638	0.819	1.002	1.184	0.661
	Observation 28 days		**0.202**	**0.311**	**0.460**	**0.631**	**0.812**	**0.994**	**1.175**	**0.655**
	Observation 56 days		0.207	0.313	0.465	0.639	0.823	1.008	1.189	0.663
	Observation 112 days		0.218	0.325	0.476	0.652	0.836	1.021	1.202	0.676
	Observation 7 days	RMSE (cm)	0.414	0.563	0.812	1.108	1.417	1.717	2.005	1.148
	Observation14 days		0.345	0.537	0.809	1.116	1.434	1.743	2.039	1.146
	Observation 28 days		**0.338**	**0.531**	**0.803**	**1.113**	**1.431**	**1.738**	**2.031**	**1.140**
	Observation 56 days		0.344	0.535	0.808	1.117	1.434	1.741	2.032	1.145
	Observation 112 days		0.362	0.550	0.824	1.138	1.458	1.768	2.061	1.166

Table 9
Comparison results of computational cost.

Methods	Flops	Training times	Inference times	Training memory	Inference memory
ConvLSTM	1.903T	0.751 s	0.253 s	22.857G	1.659G
Ours	0.302T	0.092 s	0.047 s	7.734G	1.973G

4.6. Computational cost experiments

The computational efficiency of models is crucial for practical algorithm applications. Therefore, in this experiment, we conduct a cost analysis of model computations. Relative to feedforward neural networks, the RNN tends to exhibit higher computational complexity due to additional recurrent computations at each time step, which are challenging to parallelize effectively using matrix operations. In order to intuitively demonstrate the advantages of our model in terms of computational efficiency, we choose the ConvLSTM method, which also performs well in SLA prediction performance, for comparative experiments.

Table 9 presents the total computational cost of models for predicting SLA fields over the next 7 days. The experiment is conducted in the APS area. The input SLA fields resolution is 1/4° longitude × 1/4° latitude. The computing device is based on the operating system Ubuntu 20.04 and a GPU NVIDIA 3090. Python and PyTorch are used to build our neural network model. The "Flops" refers to Floating Point Operations of the model. The "Training Time" indicates the time taken by the model to process each sample during training, while the "Inference Time" indicates the time required by the model to process each sample during inference. The "Training Memory" refers to the memory size occupied by the model's parameters when training on a single sample. The "Inference Memory" represents the memory size occupied by the model's parameters when performing inference on a single sample.

According to the Flops, the computation speed of our UNet-TSTA model has increased by 6.3 times than the advanced ConvLSTM method. Compared to ConvLSTM, the UNet-TSTA model reduces the number of parameters during training by 66.16%, while maintaining a similar number of parameters during inference.

ConvLSTM is a RNN-based variant: the output at each time step depends on the computation result from the previous time step. As a result, there is a temporal dependency during training and inference, making parallelization difficult, which significantly increases computation time. In contrast, UNet-TSTA uses the UNet and ViT architectures that support parallel processing of all time steps in the input sequence along the time dimension, thus improving computational efficiency. RNN-based variant structures tend to maintain multiple states (such as hidden states and memory units) at each time step, which makes the computational graph complex, resulting in frequent state updates and higher memory usage. UNet-TSTA, however, adopts an end-to-end approach where multiple time-step predictions are output at once, without the need to save cyclic states, resulting in less memory usage.

5. Conclusion

To address the challenge of accurately and efficiently predicting SLA fields, we have proposed a novel UNet-TSTA model based on UNet and Transformer Attention, aiming at enhancing both the accuracy and speed of SLA predictions. The parallel convolution operations of CNN improve the training and inference speed of the model, as well as its capability to process 2D spatial information. The UNet backbone architecture enhances the model's ability to capture and analyze sea eddies at different scales. The TSTA module consists of two parts: STA and TTA. The STA module utilizes the attention mechanism of P-ViT to capture longer-range spatial dependencies, thereby enhancing the model's global information perception of the SLA field. The TTA module ingeniously uses the free combination of temporal and spatial dimensions, creatively constructing two spatiotemporal planes, and further utilizes the P-ViT method to capture spatiotemporal dependencies.

We validate our approach on real SLA observational historical datasets. Compared to one of the state-of-the-art methods in the SLA prediction domain, MLFrnn, our UNet-TSTA model shows a performance lead of 17.3% in RMSE. Additionally, compared to mainstream ConvLSTM methods in recent years, our model achieves lower prediction errors and improves computational efficiency by 6.3 times. In the future, we would like to collect and utilize the multi-modal data such as sea surface height, temperature, and air pressure. The coupling relationships among these multi-modal data will facilitate further improvement in SLA field prediction performance.

CRediT authorship contribution statement

Qinxuan Wang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **Jun Bai:** Formal analysis, Writing – review & editing, Supervision, Project administration. **Yineng Li:** Resources. **Shiming Xiang:** Project administration. **Xiaoqing Chu:** Resources. **Yue Sun:** Methodology, Formal analysis. **Tielin Zhang:** Methodology, Formal analysis.

Declaration of competing interest

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

Acknowledgments

This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA0370203, and the National Natural Science Foundation of China under Grant 62306310.

References

Bonaduce, A., Pinardi, N., Oddo, P., Spada, G., Larnicol, G., 2016. Sea-level variability in the Mediterranean Sea from altimetry and tide gauges. Clim. Dyn. 47 (9), 2851–2866.

Carton, J.A., Giese, B.S., Grodsky, S.A., 2005. Sea level rise and the warming of the oceans in the simple ocean data assimilation (SODA) ocean reanalysis. J. Geophys. Res.: Ocean. 110 (C9), 1–8.

Chelton, D.B., Schlax, M.G., Lyman, J.M., Johnson, G., 2003. Equatorially trapped rossby waves in the presence of meridionally sheared baroclinic flow in the Pacific ocean. Prog. Oceanogr. 56 (2), 323–380.

Chen, J., Wilson, C., Chambers, D., Nerem, R., Tapley, B., 1998. Seasonal global water mass budget and mean sea level variations. Geophys. Res. Lett. 25 (19), 3555–3558.

Cui, H., Tang, D., Liu, H., Sui, Y., Gu, X., 2023. Composite analysis-based machine learning for prediction of tropical cyclone-induced sea surface height anomaly. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 2644–2653.

Cui, W., Yang, J., Jia, Y., Zhang, J., 2022. Oceanic eddy detection and analysis from satellite-derived SSH and SST fields in the Kuroshio extension. Remote. Sens. 14 (22), 1–18.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. pp. 1–22, arXiv preprint arXiv:2010.11929.

Dudukcu, H.V., Taskiran, M., Taskiran, Z.G.C., Yildirim, T., 2023. Temporal convolutional networks with RNN approach for chaotic time series prediction. Appl. Soft Comput. 133, 109945–109959.

Gao, R., Zhang, R., Wen, M., Li, T., 2019. Interdecadal changes in the asymmetric impacts of ENSO on wintertime rainfall over China and atmospheric circulations over western north Pacific. Clim. Dyn. 52, 7525–7536.

Gregory, J.M., Lowe, J., 2000. Predictions of global and regional sea-level rise using AOGCMs with and without flux adjustment. Geophys. Res. Lett. 27 (19), 3069–3072.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al., 2022. A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1), 87–110.

Hu, S., Li, S., Zhang, Y., Guan, C., Du, Y., Feng, M., Ando, K., Wang, F., Schiller, A., Hu, D., 2021. Observed strong subsurface marine heatwaves in the tropical western Pacific ocean. Environ. Res. Lett. 16 (10), 1–10.

Huo, F., Liu, K., Dong, H., Ren, W., Dong, S., 2024. Research on cuttings image segmentation method based on improved MultiRes-unet plus plus with attention mechanism. Signal Image Video Process. 18 (SUPPL 1, 1), 799–808.

Jiang, L., Duan, W., Wang, H., 2024. The sensitive area for targeting observations of paired mesoscale eddies associated with sea surface height anomaly forecasts. J. Geophys. Res.: Ocean. 129 (2), 1–18.

Jiang, L., Duan, W., Wang, H., Liu, H., Tao, L., 2023. Evaluation of the sensitivity on mesoscale eddy associated with the sea surface height anomaly forecasting in the Kuroshio extension. Front. Mar. Sci. 10, 1–10.

Kessler, W.S., 1990. Observations of long rossby waves in the northern tropical Pacific. J. Geophys. Res.: Ocean. 95 (C4), 5183–5217.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. ACM Comput. Surv. 54 (10s), 1–41.

Li, J., Du, J., Zhu, Y., Guo, Y., 2023. Survey of transformer-based object detection algorithms. Comput. Eng. Appl. 59 (10), 48–64.

Li, J., Un, K.-F., Yu, W.-H., Mak, P.-I., Martins, R.P., 2021. An FPGA-based energy-efficient reconfigurable convolutional neural network accelerator for object recognition applications. IEEE Trans. Circuits Syst. II: Express Briefs 68 (9), 3143–3147.

Li, G., Zhang, M., Zhang, Q., Lin, Z., 2022. Efficient binary 3D convolutional neural network and hardware accelerator. J. Real-Time Image Process. 19 (1), 61–71.

Li, R.C., Zhou, W., 2012. Changes in western Pacific tropical cyclones associated with the El Niño–southern oscillation cycle. J. Clim. 25 (17), 5864–5878.

Liao, Y., Lu, S., Yin, G., 2024. Short-term and imminent rainfall prediction model based on ConvLSTM and SmaAT-UNet. Sensors 24 (11), 3576–3590.

Ling, Y., Wang, Y., Liu, Q., Yu, J., Xu, L., Zhang, X., Liang, P., Kong, D., 2024. EPolar-UNet: An edge-attending polar UNet for automatic medical image segmentation with small datasets. Med. Phys. 51 (3), 1702–1713.

Liu, Z., Xiao, X., Li, C., Ma, S., Rangyu, D., 2022. Optimizing convolutional neural networks on multi-core vector accelerator. Parallel Comput. 112, 102945–102957.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Metzger, E.J., Hurlburt, H.E., 2001. The nondeterministic nature of Kuroshio penetration and eddy shedding in the South China Sea. J. Phys. Oceanogr. 31 (7), 1712–1732.

Meyers, G., 1979. On the annual rossby wave in the tropical north Pacific ocean. J. Phys. Oceanogr. 9 (4), 663–674.

Miles, E.R., Spillman, C.M., Church, J.A., McIntosh, P.C., 2014. Seasonal prediction of global sea level anomalies using an ocean–atmosphere dynamical model. Clim. Dyn. 43, 2131–2145.

Nan, F., He, Z., Zhou, H., Wang, D., 2011. Three long-lived anticyclonic eddies in the northern South China Sea. J. Geophys. Res.: Ocean. 116 (C5), 1–15.
Ning, P., Zhang, C., Zhang, X., Jiang, X., 2021. Short-to medium-term sea surface height prediction in the Bohai Sea using an optimized simple recurrent unit deep network. Front. Mar. Sci. 8, 1–12.
Ningsih, N.S., Hanifah, F., Tanjung, T.S., Yani, L.F., Azhar, M.A., 2020. The effect of tropical cyclone nicholas (11–20 february 2008) on sea level anomalies in Indonesian waters. J. Mar. Sci. Eng. 8 (11), 948–965.
Norris, R., Turner, S.K., Hull, P., Ridgwell, A., 2013. Marine ecosystem responses to cenozoic global change. Science 341 (6145), 492–498.
Peng, F., Deng, X., 2020a. Improving precision of high-rate altimeter sea level anomalies by removing the sea state bias and intra-1-Hz covariant error. Remote Sens. Environ. 251, 1–13.
Peng, F., Deng, X., 2020b. Validation of sentinel-3A SAR mode sea level anomalies around the Australian coastal region. Remote Sens. Environ. 237, 1–16.
Peng, F., Deng, X., Cheng, X., 2021. Quantifying the precision of retracked Jason-2 sea level data in the 0–5 km Australian coastal zone. Remote Sens. Environ. 263, 1–16.
Peng, F., Deng, X., Shen, Y., 2024. Assessment of sentinel-6 SAR mode and reprocessed Jason-3 sea level measurements over global coastal oceans. Remote Sens. Environ. 311, 1–17.
Popescu, M.-C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N., 2009. Multilayer perceptron and neural networks. WSEAS Trans. Circuits Syst. 8 (7), 579–588.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI) 234–241.
Song, T., Jiang, J., Li, W., Xu, D., 2020. A deep learning method with merged LSTM neural networks for SSHA prediction. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 2853–2860.
Su, J., Byeon, W., Kossaifi, J., Huang, F., Kautz, J., Anandkumar, A., 2020. Convolutional tensor-train LSTM for spatio-temporal learning. Adv. Neural Inf. Process. Syst. 33, 13714–13726.
Trebing, K., Stanczyk, T., Mehrkanoon, S., 2021. SmaAtUNet: Precipitation nowcasting using a small attention-UNet architecture. Pattern Recognit. Lett. 145, 178–186.
Wang, G., Wang, X., Wu, X., Liu, K., Qi, Y., Sun, C., Fu, H., 2022. A hybrid multivariate deep learning network for multistep ahead sea level anomaly forecasting. J. Atmos. Ocean. Technol. 39 (3), 285–301.
Whan, K., Alexander, L., Imielska, A., McGree, S., Jones, D., Ene, E., Finaulahi, S., Inape, K., Jacklick, L., Kumar, R., et al., 2014. Trends and variability of temperature extremes in the tropical western Pacific. Int. J. Climatol. 34 (8), 2585–2603.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.
Wu, H., Yao, Z., Wang, J., Long, M., 2021. MotionRNN: A flexible model for video prediction with spacetime-varying motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15435–15444.
Yang, X., Dang, Z., Yu, J., Zhong, Z., Chang, M., Zhang, Z., 2024. Sequence recommendation algorithm fusing filter and transformer under joint training. J. Intell. Fuzzy Systems 46 (1), 941–953.
Yoshida, S., Morimoto, T., Ushio, T., Kawasaki, Z., 2007. ENSO and convective activities in southeast Asia and western Pacific. Geophys. Res. Lett. 34 (21), 1–6.
You, T., Wu, R., Liu, G., Chai, Z., 2021. Contribution of precipitation events with different consecutive days to rainfall change over Asia during ENSO years. Theor. Appl. Climatol. 144, 147–161.
Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. pp. 1–8, arXiv preprint arXiv:1409.2329.
Zhao, Z., Liu, B., Li, X., 2014. Internal solitary waves in the China seas observed using satellite remote-sensing techniques: a review and perspectives. Int. J. Remote Sens. 35 (11–12), 3926–3946.
Zhao, X., Yuan, D., Wang, J., 2023. Sea level anomalies in the southeastern tropical Indian ocean as a potential predictor of la Niña beyond one-year lead. Front. Mar. Sci. 10, 1–18.
Zheng, Q., Hu, J., Zhu, B., Feng, Y., Jo, Y.-H., Sun, Z., Zhu, J., Lin, H., Li, J., Xu, Y., 2014. Standing wave modes observed in the South China Sea deep basin. J. Geophys. Res.: Ocean. 119 (7), 4185–4199.
Zhou, Y., Lu, C., Chen, K., Li, X., 2021. Multilayer fusion recurrent neural network for sea surface height anomaly field prediction. IEEE Trans. Geosci. Remote Sens. 60, 1–11.
Zhou, Y., Ren, T., Chen, K., Gao, L., Li, X., 2024. Graph-based memory recall recurrent neural network for mid-term sea surface height anomaly forecasting. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17, 6642–6657.