Long-Term Prediction of Sea Surface Temperature by Temporal Embedding Transformer With Attention Distilling and Partial Stacked Connection

Hao Dai, Zhigang He, Guomei Wei, Famei Lei, Xining Zhang, Weijie Zhang, and Shaoping Shang

***Abstract*—Sea surface temperature (SST) is one of the most important parameters in the global ocean–atmosphere system, and its long-term changes will have a significant impact on global climate and ecosystems. Accurate prediction of SST, therefore, especially the improvement of long-term predictive skills is of great significance for fishery farming, marine ecological protection, and planning of maritime activities. Since the effective and precise description of the long-range dependence between input and output requires higher model prediction ability, it is an extremely challenging task to achieve accurate long-term prediction of SST. Inspired by the successful application of the transformer and its variants in natural language processing similar to time-series prediction, we introduce it to the SST prediction in the China Sea. The model Transformer with temporal embedding, attention Distilling, and Stacked connection in Part (TransDtSt-Part) is developed by embedding the temporal information in the classic transformer, combining attention distillation and partial stacked connection, and performing generative decoding. High-resolution satellite-derived data from the National Oceanic and Atmospheric Administration is utilized, and long-term SST predictions with day granularity are achieved under univariate and multivariate patterns. With root mean square error and mean absolute error as metrics, the TransDtSt-Part outperforms all competitive baselines in five oceans (i.e., subareas of Bohai, Yellow Sea, East China Sea, Taiwan Strait, and South China Sea) and six prediction horizons (i.e., 30, 60, 90, 180, 270, and 360 days). Experimental results demonstrate that the performance of the innovative model is encouraging and promising for the long-term prediction of SST.**

***Index Terms*—Attention distilling, China Sea, long-term prediction, partial stacked connection, sea surface temperature (SST), temporal transformer.**

I. Introduction

SEA surface temperature (SST) provides basic information about the global climate system and is an important parameter for weather prediction and atmospheric model simulations [1], [2]. SST measurements benefit a wide range of operational applications, including climate monitoring, fishery farming, maritime commercial activities, etc. From north to south, the China Sea is mainly composed of the Bohai Sea, the Yellow Sea, the East China Sea, the Taiwan Strait, and the South China Sea. The China Sea spans four climatic zones: temperate zone, warm temperate zone, subtropical zone, and tropical zone. The confluence of ocean currents and the development of fronts have created many important fishing grounds, and the marine aquaculture industry is developed. Meanwhile, red tide disasters are more serious in spring and summer every year. Moreover, the China Sea is part of the Maritime Silk Road and also has one of the busiest shipping lanes in the world. Hence, accurate prediction of SST in the China Sea is crucial for an in-depth understanding of marine fishery farming, ecological change dynamics, and maritime activity planning, which are very important to the production and lives of Chinese people.

SST is mainly affected by many factors, such as solar radiation, air–sea heat flux, and diurnal winds, which form a complex and changeable vertical structure that changes over time. Changes in solar radiation affect the energy balance at the sea surface. Variations in cloud cover, atmospheric conditions, and the time of day influence the radiation balance, impacting SST. Positive heat flux, where the ocean gains heat from the atmosphere, leads to an increase in SST, and negative heat flux results in a decrease. Diurnal winds can affect the vertical mixing of the ocean layers. During the day, solar heating can lead to the development of sea breezes, causing mixing and influencing SST. At night, cooling processes may dominate. Due to the irregularity of thermal radiation, flux, and the uncertainty of wind blowing over the sea surface, it is difficult to construct effective and reliable mathematical equations to describe the causal relationship between SST and these factors, resulting in many difficulties in accurately predicting SST [3], [4].

Currently, SST prediction methods are mainly divided into three categories: numerical models based on physics, data-driven techniques, and hybrids of the two. The numerical models rely on the dynamical and thermal equations based on the required initial and boundary conditions. They describe the physical states using partial differential equations and make predictions of future SST after conducting a large number of calculations to derive numerical solutions [5], [6], [7]. Since no

Manuscript received 7 October 2023; revised 19 December 2023; accepted 17 January 2024. Date of publication 25 January 2024; date of current version 12 February 2024. This work was supported by the Fujian Province Marine Economic Development Subsidy Fund Project under Grant ZHHY-2019-2. *(Corresponding authors: Hao Dai; Shaoping Shang.)*

Hao Dai, Zhigang He, Guomei Wei, Famei Lei, Weijie Zhang, and Shaoping Shang are with the Institute of Ocean Exploration Technology, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361005, China (e-mail: daihaozxn@163.com; zghe@xmu.edu.cn; weiguomei@163.com; lfm101659@163.com; oot@xmu.edu.cn; spshang@xmu.edu.cn).

Xining Zhang is with the Fujian Key Laboratory of Light Propagation and Transformation, College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China (e-mail: zhangxn_hqu@163.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2024.3357191, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2024.3357191

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

clear physical explanation can be given for the mechanism of SST generation and evolution, the construction of such models is generally inaccurate, relatively complex, and computationally expensive. They are more suitable for large-scale SST prediction with rough resolution. Starting from the characteristics of the data and internal laws, the data-driven methods learn rules from the data and make predictions by training a large number of known samples. The methods mainly include a statistical approach [8], genetic algorithm [9], and deep learning [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. Under the premise that a large amount of reliable observation data is available, this type of method can obtain satisfactory prediction results quickly. Now the hybrid approach combines the numerical model and artificial neural network for SST prediction [4]. Due to the use of two models, such a method is the most intricate, consuming the most computing resources and taking the longest time.

Through literature research, we find that most studies on SST prediction using daily average data only focus on short- and medium-term prediction (lead time ≤ 10 days) probably because it is difficult to find reliable dependencies from the long-term complex time patterns in the future [10], [11], [12], [14], [15], [16], [18], [19], [21], [22], [23]. The so-called long-term prediction is weekly mean or monthly mean as the granularity interval which is too rough [13], [17], [20], [24], [25]. No long-term prediction studies with a prediction horizon $\geq$30 days based on daily average SST have been reported.

Extending the lead time of SST is a key requirement for practical applications, such as marine ecosystems and long-term planning of marine activities. Benefiting from the self-attention mechanism, transformer has gained a huge advantage in modeling sequential data dependencies, such as natural language processing (NLP, [26]) and audio processing [27]. This brings light to its introduction for SST prediction. However, the prediction task is extremely challenging in the long-term prediction horizon setting. First, it is unreliable to discover temporal dependencies directly from long-term time series because dependencies may be masked by entangled temporal patterns. Second, due to the quadratic complexity of the sequence length, the canonical transformer with the self-attention mechanism is computationally prohibitive and difficult to be directly used for long-term prediction.

Based on the traditional transformer, therefore, this article uses generative decoding, embeds timing information, and adds attention distillation/partial stacked connection to construct the model named **Trans**former with temporal embedding, attention **Dist**illing, and **St**acked connection in **Part** (TransDtSt-Part). Five typical oceans of the China Sea are selected. With the help of daily average SST, using univariate and multivariate prediction patterns, the long-term prediction skill of TransDtSt-Part is comprehensively evaluated by comparing it with multiple baseline models.

The rest of this article is organized as follows. The details of the model proposed in this article are provided in Section II. Section III clarifies the data used in this article and the research area. Section IV evaluates the proposed model via experiments. Finally, Section V concludes this article.

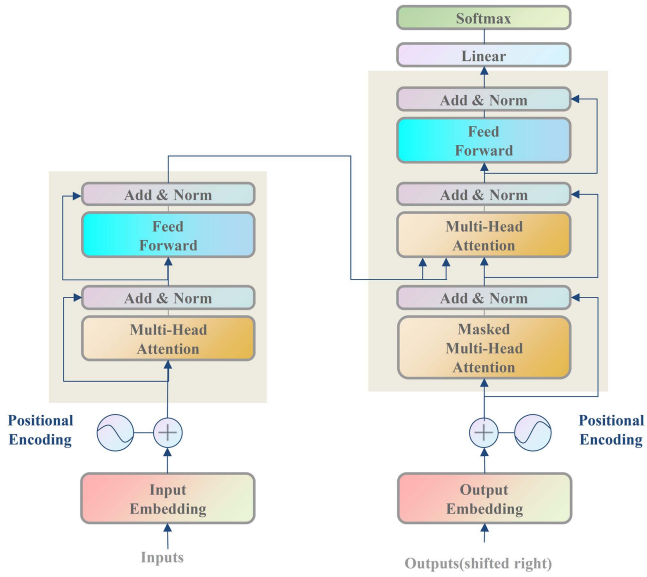

Fig. 1. Classic transformer architecture.

II. Methods

A. Classic Transformer

Similar to recurrent neural networks (RNN), the classic transformer [28] is designed to process sequential input data and is applied to tasks, such as translation and text summarization. It features self-attention, which differentially weights the importance of each part of the input (including the recursive output). Unlike RNNs, transformers process all inputs at once during training. The self-attention mechanism provides context for any position in the input sequence. Hence, to a large extent, transformer solves the problems of RNN training inefficiency and long-range dependency insufficiency. It should be noted that during inference, the transformer decoder still uses the autoregressive method for dynamic decoding.

The classic transformer for NLP first parses the input text into tokens through a byte-pair-encoded tokenizer, and each token is converted into a vector through word embedding. Then, the tagged position information is added to the word embedding.

As shown in Fig. 1, the transformer model uses an encoder and decoder architecture. The encoder consists of encoding layers that iteratively process the input layer by layer, while the decoder consists of decoding layers that perform the same operation on the output of the encoder.

The function of each encoder layer is to generate an encoding containing information about which parts of the input are related to each other. It passes its encoding as input to the next encoder layer. Each decoder layer does the opposite, taking all the encodings and using their combined contextual information to generate an output sequence. To achieve this, each encoder and decoder layer uses a multihead scaled dot-product attention mechanism. For each part of the input, attention weighs the relevance of all other parts and extracts information from them to produce output. Each decoder layer has an additional cross-attention for incorporating the output of the encoder. Both the

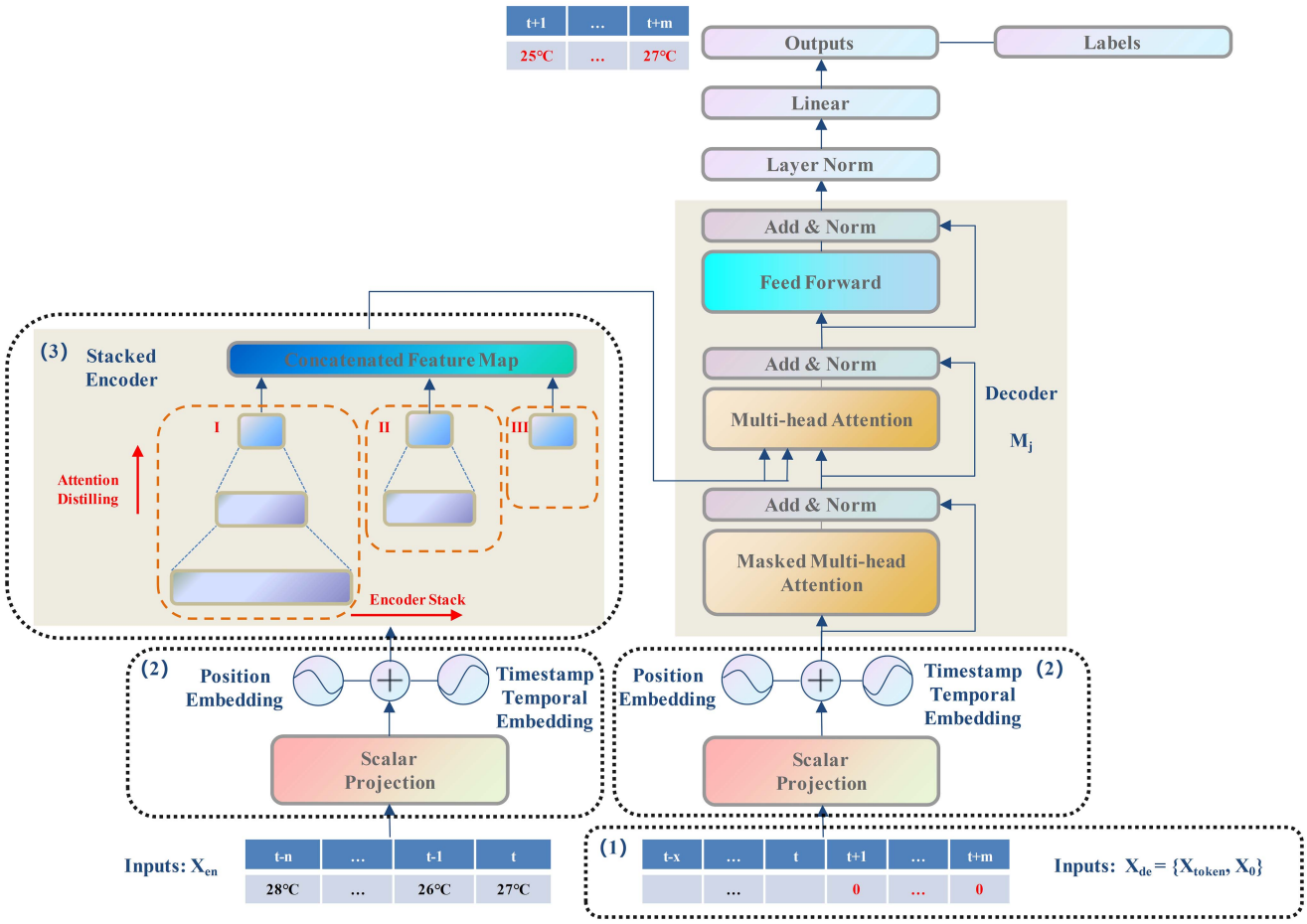

Fig. 2. Structure of TransDtSt-Part.

encoder and decoder layers have a feedforward neural network for additional processing of the output and contain residual connection and layer normalization steps.

B. TransDtSt-Part Model

LSTM and its variants, which have been used in NLP, have also been employed to predict SST [23], [24]. Inspired by the great success of transformer and the similarity to a certain extent between SST prediction and sequence tasks, such as machine translation, according to the characteristics of SST time-series prediction, we make the following improvements to construct the model TransDtSt-Part.

1) Employing generative decoding to improve the recursive output form of the traditional transformer, which is inefficient and has accumulated errors, to heighten the prediction skill.
2) Considering the time-dimensional information that has not been employed in the previous SST prediction works, and adding timestamp embedding.
3) Performing attention distilling and partial stacked connection of the encoder to improve the prediction accuracy and enhance the robustness.

The article designs a deep-learning network for SST long-term prediction and the structure of the TransDtSt-Part with attention distilling and partial stacked connection is exhibited in Fig. 2.

From bottom to top, these three improvements correspond to the black dotted boxes marked (1), (2), and (3) in Fig. 2, respectively. These improvements will be detailed as follows.

1) Generative Decoding: As shown in Fig. 2, the input X_{en} of the stacked encoder before embedding is the daily average data of SST for n days from the historical $(t-n)$th day to the current tth day (seq_{len}). The start token strategy is successfully applied to NLP [26], and we extend it into a generative way in the article. Instead of choosing specific flags as the token, we sample a long sequence in the input, that is, sampling a sequence of length ($\text{label}_{\text{len}}$) in the input sequence of the stacked encoder before embedding (namely, the daily average data of SST from the day $(t-x)$th to the day tth in Fig. 2, $x \leq n$, i.e., $\text{label}_{\text{len}} \leq \text{seq}_{\text{len}}$). The initial values of the expected prediction horizons [day $(t+1)$th–day $(t+m)$th, a total of pred_{len} days] are m zeros, and are connected as the input of the decoder before embedding.

In Fig. 2, the SST sequence X_{token} is used as the start token to lead the initial values of the target sequence into the classic transformer decoder. Through a forward process, the multistep

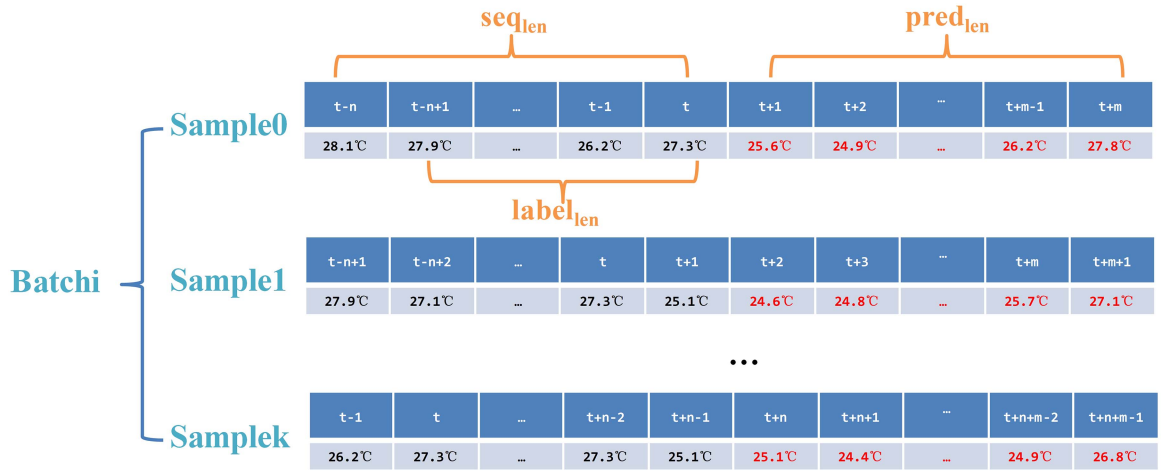

Fig. 3. Sample formation instructions.

prediction output (namely, the "outputs" in Fig. 2) can be obtained. In this way, the sudden drop in the inference speed of the original "dynamic decoding" in the long prediction is alleviated, and the accumulation of errors is avoided.

As shown in Fig. 3, for a new sample, the lookback window SST with a length of seq_{len} moves forward one step as a whole, and the pred_{len} prediction also moves one step forward accordingly. Repeat this to form k samples and one batch.

2) Temporal Embedding: In previous articles on deep learning to predict SST, temporal dimension information has not been exploited. The ability to achieve remote independence requires global information, for example, hierarchical timestamps (week, month, and year). These are rarely exploited in canonical self-attention, so a query-key mismatch between the encoder and decoder may lead to a potential drop in prediction performance.

In this article, before the SST enters the stacked encoder and decoder, we perform three forms of embedding on the input and superimpose them. The three forms of embedding are as follows.

1) *Position embedding:* Similar to NLP, it is required to deal with longer inputs in long-term SST prediction, so a parallel input strategy is adopted. However, considering the contextual relationship between time-series SST data, position embedding needs to be added. Hence, we follow the embedding operation in [28]. Specifically, it is formalized by

$$\begin{aligned} \text{PE}_{(\text{pos},2i)} &= \sin\left(\text{pos}/\left((10\,000)^{\frac{2i}{d_{\text{model}}}}\right)\right) \\ \text{PE}_{(\text{pos},2i+1)} &= \cos\left(\text{pos}/\left((10\,000)^{\frac{2i}{d_{\text{model}}}}\right)\right) \end{aligned} \tag{1}$$

where pos represents position, i represents dimension, and d_{model} represents embedding vector dimension.

2) *Timestamp embedding:* Temporal embedding is performed on the timestamp information corresponding to the encoder and decoder inputs, respectively, to access the global context information. Embedded coding is performed according to the timestamp interval type. This article uses daily average SST data, that is, the timestamp interval is "day." Assuming that each day is indexed from 0, it is encoded into week, month, and year. The encoding formulas are as follows:

$$\text{indexdayofweek}_{\text{encoded}} = \frac{2 \times \text{indexdayofweek}}{6} - 1 \tag{2}$$

$$\text{indexdayofmonth}_{\text{encoded}} = \frac{2 \times \text{indexdayofmonth}}{30} - 1 \tag{3}$$

$$\text{indexdayofyear}_{\text{encoded}} = \frac{2 \times \text{indexdayofyear}}{365} - 1 \tag{4}$$

where $\text{indexdayofweek} \in [0, \ldots, 6]$, $\text{indexdayofmonth} \in [0, \ldots, 30]$, and $\text{indexdayofyear} \in [0, \ldots, 365]$.

With the help of (2), (3), and (4), the timestamp "day" is encoded into three vectors and then passes through a linear layer with the input dimension 3 and output dimension d_{model} for timestamp embedding.

3) *Scalar projection:* To align the dimension, one-dimensional (1-D) convolution is performed with kernelsize 3, stride 1, padding 1, and the circular padding mode, and the encoder and decoder inputs are separately projected.

3) Attention Distilling and Partial Stacked Connection: In the long-term prediction task of SST, more computing resources are consumed because of long time series. To improve the prediction ability of the model, given the redundant combination of values V in the feature map of the entire encoder, we first use the distilling operation to sample, greatly reducing the input size, and prioritize the attention scores with dominant features. Then, the input is halved encoder-by-encoder to enhance the distillation robustness, and the distilling layer is reduced accordingly

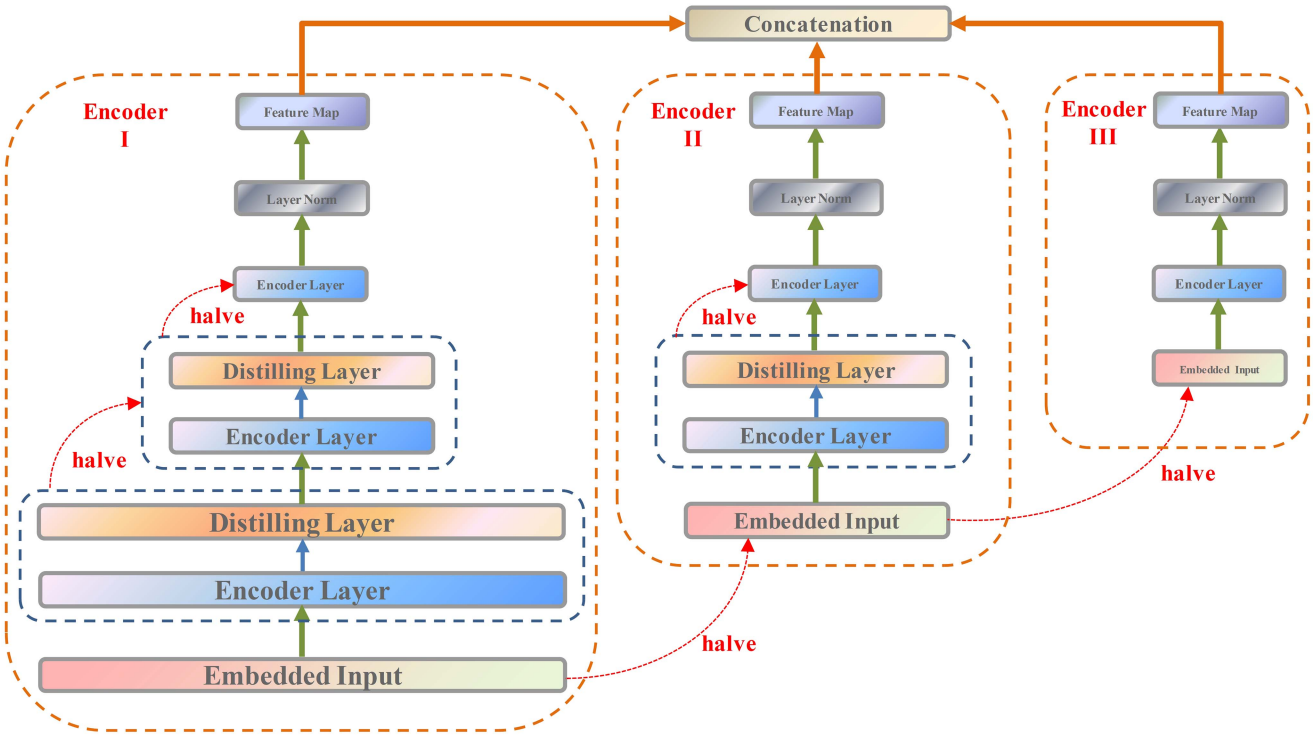

Fig. 4. Encoder stacking process in TransDtSt-Part network.

to align the output of each encoder. Finally, the outputs of all or partial encoders are concatenated to form the final feature map.

The whole process of the encoder stacking is shown in Fig. 4 (take stacking three encoders as an example).

According to the encoder settings of the traditional transformer, Fig. 5(a) exhibits that the encoder layer mainly includes scaling dot product multihead self-attention, residual connection, layer normalization, feedforward, activation, and dropout, where feedforward mainly includes the 1-D convolution Conv1d with parameters kernelsize 1, padding 0, and stride 1 to realize the effect of a fully connected structure [28], activation, and dropout.

As shown in Fig. 5(b), the distilling layer mainly consists of downsampling (1-D convolution, parameters: convolution kernelsize 3, padding 2, stride 1, padding mode circular), batch normalization, activation (function selected as ELU [29]), and maximum pooling. In particular, stride $= 2$ in the maximum pooling is employed to achieve halved self-attention "distilling."

It should be pointed out that we investigate the effect of the encoder stacking connection length on the SST prediction performance, and discover that probably due to receiving more long-term information, a longer stack is more sensitive to the input, resulting in the prediction effect of connecting all encoders is inferior to that of partial connecting encoders (for details, please refer to the prediction indicators of models TransDtSt-All and TransDtSt-Part in Supplementary Material A. Ablation study). Therefore, in the model TransDtSt-Part, we take the most robust strategy of joining Encoder I and Encoder III of Fig. 4.

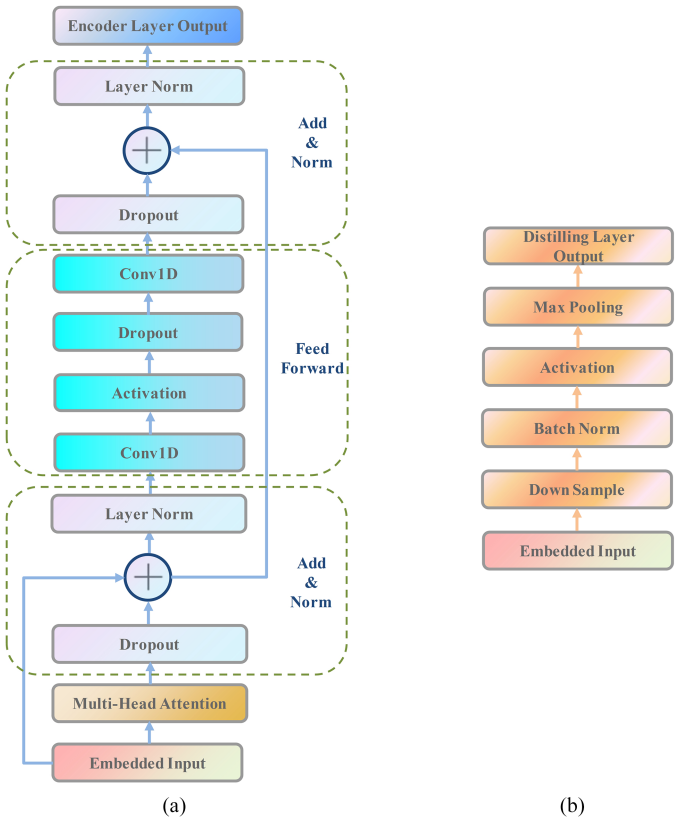

Fig. 5. Encoder layer and distilling layer in stacked encoder. (a) Encoder layer. (b) Distilling layer.

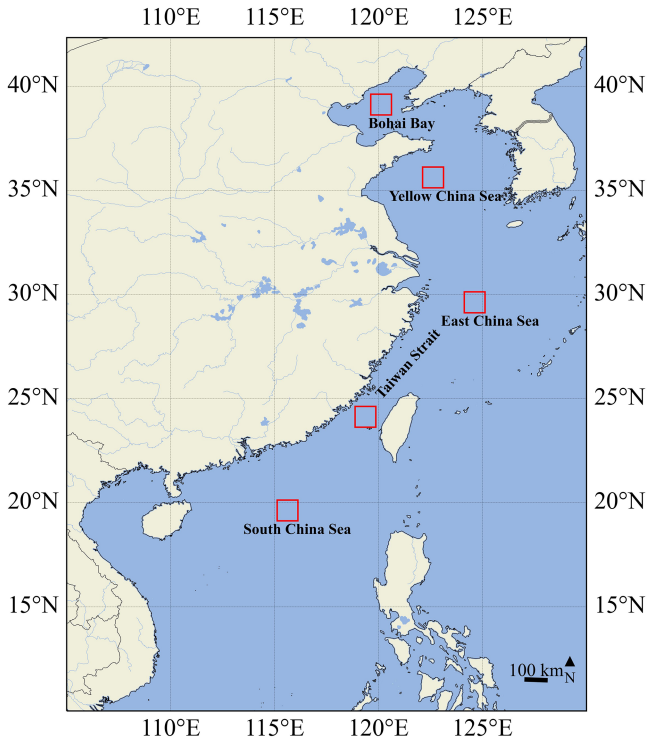

Fig. 6. Research area.

III. Data and Research Area

A. Data

Similar to many works of literature that employ deep-learning techniques to predict SST [19], [20], [21], [22], the source data come from multiyear daily average data in the Optimum Interpolation High-Resolution SST Dataset Version 2 (OISST) provided by the Physical Sciences Laboratory of National Oceanic and Atmospheric Administration. The time range of the dataset is from September 1981 to the present, and the spatial coverage is 89.875 °S to 89.875 °N and 0.125 °E to 359.875 °E with a spatial resolution of 0.25 ° latitude by 0.25 ° longitude. Models are trained, validated, and tested on OISST as ground truth.

B. Research Area

To accurately compare the effects of our method in different China Sea areas, the interest subareas of the five typical oceans should be equal in size. Since the Taiwan Strait is a strip-shaped region and the research area that can be selected is relatively limited, 5×5 pixels is basically the largest square region in the OISST dataset. Meanwhile, we also consider the application effect of our model in nearshore and offshore areas. The subregions of the Taiwan Strait and Bohai Sea include the coastal areas, while those in the Yellow Sea, East China Sea, and South China Sea belong to the open sea. As shown in the red boxes of Fig. 6, the longitude and latitude coordinate ranges of the five sea areas from north to south are Bohai Sea (119.625 °E–120.625 °E, 38.625 °N–39.625 °N), Yellow Sea (122.125 °E–123.125 °E, 35.125 °N–36.125 °N), East China Sea (124.125 °E–125.125 °E, 29.125 °N–30.125 °N), Taiwan Strait (118.875 °E–119.875 °E, 23.625 °N–24.625 °N), and South China Sea (115.125 °E–116.125 °E, 19.125 °N–20.125 °N).

IV. Experiments and Results

A. Dataset Partitioning and Preprocessing Method

We use a total of 40 years of data from 1982 to 2021 and divide by the volume of 7:1:2, i.e., 1982–2009 as the training dataset, 2010–2013 as the validation set for hyperparameters tuning, and 2014–2021 as a test set to evaluate the generalization ability of the model in the face of new data. In this article, mean-variance normalization is employed to preprocess the SST training set and applied to the validation and test datasets in the same way.

B. Baseline Models

Four baseline models are selected for predictive performance comparison, i.e., RNN-based model LSTM [30], CNN-based model TCN [31], transformer-based model informer [32], and interpretable time-series prediction model N-BEATS [33].

C. Hardware Platform and Software Environment

The experiments of the proposed model and baseline models are carried out on the following hardware configurations: CPU-Intel i9-9900k, RAM-64G, NVIDIA Geforce RTX 2080Ti 11G. The software environment adopts the Win10 Operating System, Integrated Development Environment Pycharm, and deep-learning framework PyTorch (1.10.1).

D. Hyperparameters

The hyperparameters to be determined in the prediction of the model TransDtSt-Part are as follows:

1) Number of encoder layers: $e_{\text{layers}} = [3, 4, 5, 6]$.
2) Number of decoder layers: $d_{\text{layers}} = [1, 2]$.
3) Generate dimensions of all sublayers and embedding layers in the model: $d_{\text{model}} = [128, 256, 512]$.
4) Dimension of feedforward network layer: $d_{\text{ff}} = [512, 1024]$ and $d_{\text{ff}} > d_{\text{model}}$.
5) Training epoch: epoch $= [10, 15, 20]$.
6) Training with the early stop strategy, involving waiting for patience: patience $= [3, 5]$.

Since the TransDtSt-Part model has a large number of hyperparameter combinations, the screening workload is relatively large. Considering the cost of time and computing resources, for some lead time in a certain ocean and univariate/multivariate prediction pattern, we first fix the parameters epoch $= 10$, patience $= 3$, and use grid search to determine the best parameters of e_{layers}, d_{layers}, d_{model}, and d_{ff}. That is, construct networks through various hyperparameter combinations within the screening range, model fitting on the training dataset, and choose the hyperparameters corresponding to the best prediction results on the validation set as the optimal values. Then, fix the selected best hyperparameters (e_{layers}, d_{layers}, d_{model}, and d_{ff}), set different epoch/patience combinations, compare the prediction performances on the validation dataset, and determine

the best epoch and patience (taking the Bohai Sea as an example, Section B in the Supplementary Material shows the optimal hyperparameter determination process and results).

Other fixed hyperparameters: the activation function is gelu, the initial learning rate lr = 0.0001, the number of heads *n*heads = 8, batchsize = 32, and dropout = 0.05.

E. Loss Function and Optimizer

We choose MSE as the loss function when predicting SST, and the loss is passed back to the whole model from the decoder output. The optimizer selects Adam.

F. Predictive Pattern

Specifically, the **univariate** prediction pattern refers to taking the historical data of the geographical center grid point of the interesting region as input, feeding each model, and obtaining the multistep prediction of the point at one time. The **multivariate** prediction pattern refers to using all the historical data of the 25 grid points of the interesting area as input, feeding each model, and getting multistep predictions for all points at once.

G. Metrics

As shown in Fig. 3, the metrics of whole length pred_{len} predictions in all samples are calculated to evaluate the predictive skill.

According to (5) and (6), the RMSE and MAE between the ground truth and the prediction are calculated as evaluation indicators. A lower RMSE or MAE indicates a better prediction

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{5}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{6}$$

where $y_i (i = 1, \ldots, n)$ is the prediction, $x_i (i = 1, \ldots, n)$is the ground truth, and n is the number of samples.

H. Main Results

Table I shows univariate SST predictive skill for five sea subregions and lead times of $\text{pred}_{\text{len}} \in \{30, 60, 90, 180, 270, \textit{and } 360\}$ days. Also, Table II is for the multivariate case. Since the change cycle of SST is in years, $\text{seq}_{\text{len}} = 360$ days is chosen. And $\text{pred}_{\text{len}} \leq \text{label}_{\text{len}} \leq \text{seq}_{\text{len}}$, $\text{seq}_{\text{len}} = \text{label}_{\text{len}} + \text{pred}_{\text{len}}$ could generally achieve better prediction results (Section C in Supplementary Material will give the prediction performance with fixed seq_{len} and different $\text{label}_{\text{len}}$). Hence, when $\text{pred}_{\text{len}} \in \{30, 60, 90, 180\}$ days, seq_{len} is equal to 360 days, correspondingly $\text{label}_{\text{len}} \in \{330, 300, 270, 180\}$ days. And when $\text{pred}_{\text{len}} \in \{270, 360\}$ days, $\text{seq}_{\text{len}} = \text{label}_{\text{len}} = 360$ days.

In addition, we examine the relationship between encoder input length and model performance. So Tables I and II also include the prediction errors of each model when $\text{seq}_{\text{len}} \in \{360, 540, 720\}$ days, and $\text{label}_{\text{len}} = \text{pred}_{\text{len}} = 360$ days.

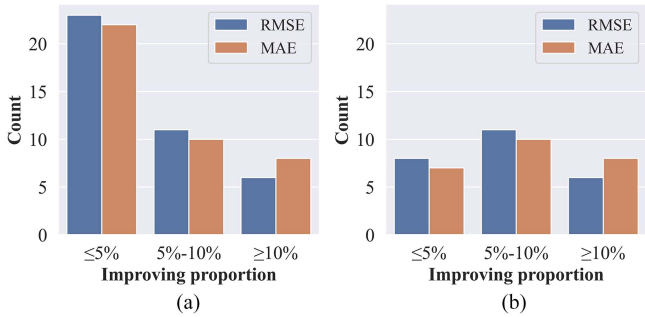

Fig. 7. **Univariate**, all oceans, number of prediction improvement rate intervals of TransDtSt-Part relative to N-BEATS. (a) All lead times. (b) Lead times ≥ 180 days.

In the prediction effect comparison of the baseline models, we call the models LSTM, N-BEATS, and TCN of the Library Darts and use the Library Optuna for hyperparameter screening. Note that N-BEATS achieves multivariate prediction by flattening the model input into a 1-D series and reshaping the output into a tensor of appropriate size, while TCN requires seq_{len} to be greater than pred_{len} during prediction, so the symbol "×" is employed to fill in the prediction blanks of $\text{seq}_{\text{len}} = \text{pred}_{\text{len}} = 360$ days in Tables I and II.

1) Analysis of SST Univariate Prediction Results: Following conditions can be found in Table I.

1) From the perspective of the lead times of 30–360 days in each ocean, the prediction error of the model TransDtSt-Part, which uses generative decoding to predict multiple values in one step, has generally resisted the extended lead time, showing the characteristics of a steady and slow rise with the prediction horizon increasing.
2) For all oceans and all lead times, the prediction error of TransDtSt-Part is smaller than other baseline models from the statistical values of the optimal prediction results (i.e., the Count row). It is verified that the long-term SST predictive skill of TransDtSt-Part is satisfactory with the day as the fine-grained interval. When compared with N-BEATS which has closer predictive performance, according to the prediction improvement rate (RMSE and MAE ≤5%, 5%–10%, and ≥10%) of TransDtSt-Part relative to N-BEATS, we count the counts of all lead times for all research subareas (a total of 40 counting points), as shown in Fig. 7(a).

   As can be seen in Fig. 7(a), about half of the predictors improve below 5%. However, if we consider the statistical results of the prediction horizons ≥180 days [that is, Fig. 7(b), a total of 25 count points], we find that the counts whose metrics improve below 5% are greatly decreased, while the number of improvements more than 5% does not change. This phenomenon reveals that compared with N-BEATS, a relatively large improvement occurs at a longer lead time. In other words, TransDtSt-Part improves more for a longer prediction horizon.
3) The prediction errors in the open seas of the Yellow Sea, the East China Sea, and the South China Sea are

TABLE I
SST UNIVARIATE PREDICTIVE SKILL OF THE MODELS WITH DIFFERENT PREDICTION HORIZONS IN THE FIVE SEA AREAS

Models / Metrics		TransDtSt-Part		LSTM		N-BEATS		TCN		Informer	
		RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)
Bohai, China	30	**0.957**	**0.747**	3.144	2.513	1.006	0.780	4.385	3.892	1.037	0.810
	60	**1.052**	**0.820**	5.386	4.218	1.087	0.857	7.799	6.897	1.155	0.917
	90	**1.170**	**0.917**	7.632	5.903	1.179	0.937	11.463	10.294	1.181	0.940
	180	**1.160**	**0.912**	12.088	9.939	1.202	0.961	16.166	14.551	1.233	0.989
	270	**1.123**	**0.884**	13.373	11.315	1.379	1.104	11.484	10.329	1.195	0.950
	360	**1.142**	**0.900**	11.748	9.606	1.288	1.023	×	×	1.195	0.942
	360[a]	**1.100**	**0.854**	8.492	7.522	1.243	0.985	10.153	7.586	1.194	0.941
	360[b]	**1.082**	**0.849**	11.785	9.658	1.372	1.092	1.476	1.178	1.184	0.934
Yellow Sea, China	30	**0.876**	**0.658**	2.522	2.028	0.880	0.663	3.384	2.977	0.950	0.714
	60	**0.981**	**0.742**	4.179	3.279	1.000	0.771	5.919	5.375	1.059	0.805
	90	**1.000**	**0.751**	5.776	4.533	1.022	0.785	8.805	7.862	1.105	0.846
	180	**1.055**	**0.802**	8.996	7.381	1.154	0.896	11.846	10.617	1.152	0.882
	270	**1.070**	**0.828**	9.521	7.958	1.161	0.905	8.658	7.711	1.173	0.919
	360	**1.068**	**0.816**	9.297	7.584	1.204	0.936	×	×	1.152	0.888
	360[a]	**1.036**	**0.787**	6.619	5.787	1.144	0.891	9.126	8.362	1.102	0.852
	360[b]	**1.004**	**0.760**	9.221	7.430	1.157	0.886	1.292	0.989	1.114	0.855
East China Sea	30	**0.793**	**0.602**	2.094	1.644	0.798	0.613	3.076	2.704	0.821	0.628
	60	**0.868**	**0.665**	3.496	2.708	0.880	0.687	5.229	4.692	0.885	0.684
	90	**0.893**	**0.692**	4.690	3.596	0.902	0.703	7.061	6.303	0.916	0.713
	180	**0.937**	**0.738**	7.519	6.139	1.016	0.801	9.249	8.340	0.976	0.759
	270	**0.930**	**0.722**	8.146	6.889	0.946	0.744	7.097	6.387	0.957	0.749
	360	**0.935**	**0.732**	7.457	6.078	0.958	0.750	×	×	0.975	0.765
	360[a]	**0.943**	**0.739**	5.119	4.499	0.963	0.751	4.072	3.587	0.959	0.748
	360[b]	**0.920**	**0.718**	7.458	6.102	0.960	0.753	1.096	0.853	0.926	0.719
Taiwan Strait	30	**1.088**	**0.833**	1.773	1.364	1.094	0.842	2.199	1.790	1.109	0.857
	60	**1.109**	**0.864**	2.600	2.056	1.132	0.879	3.611	3.099	1.133	0.881
	90	**1.138**	**0.882**	3.285	2.558	1.146	0.884	4.597	4.061	1.152	0.889
	180	**1.122**	**0.864**	4.950	3.968	1.217	0.951	6.293	5.454	1.156	0.901
	270	**1.160**	**0.904**	5.046	4.196	1.175	0.909	4.799	4.228	1.190	0.921
	360	**1.135**	**0.881**	4.716	3.772	1.203	0.934	×	×	1.159	0.897
	360[a]	**1.154**	**0.896**	3.422	2.914	1.237	0.955	2.257	1.666	1.181	0.903
	360[b]	**1.161**	**0.906**	4.888	3.931	1.212	0.939	1.302	0.997	1.200	0.922
South China Sea	30	**0.754**	**0.599**	1.394	1.101	0.773	0.610	1.543	1.235	0.764	0.611
	60	**0.805**	**0.647**	1.886	1.465	0.835	0.666	2.467	2.019	0.831	0.671
	90	**0.817**	**0.655**	2.501	1.915	0.836	0.663	3.430	2.919	0.873	0.706
	180	**0.862**	**0.696**	3.376	2.751	0.907	0.719	4.408	3.985	0.911	0.741
	270	**0.850**	**0.686**	3.883	3.220	0.916	0.732	2.860	2.419	0.893	0.715
	360	**0.856**	**0.692**	3.486	2.802	0.944	0.761	×	×	0.898	0.723
	360[a]	**0.884**	**0.713**	2.459	2.161	0.940	0.754	1.112	0.881	0.923	0.741
	360[b]	**0.895**	**0.708**	3.293	2.659	0.951	0.761	1.001	0.786	0.920	0.728

[a] means seq_{len}=540 days, $label_{len}$=360 days, and $pred_{len}$=360 days while [b] means seq_{len} =720 days, $label_{len}$ =360 days, and $pred_{len}$ =360 days.
Bold values highlights the optimal prediction results for each lead time in every sea area.
The "Count" row represents the count of the optimal value of each model under all lead times in all sea areas.

TABLE II
SST **MULTIVARIATE** PREDICTIVE SKILL OF THE MODELS WITH DIFFERENT PREDICTION HORIZONS IN THE FIVE SEA AREAS

Models / Metrics		TransDtSt-Part		LSTM		N-BEATS		TCN		Informer	
		RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)	RMSE (℃)	MAE (℃)
Bohai, China	30	**1.029**	**0.800**	3.019	2.377	1.059	0.824	4.371	3.875	1.131	0.879
	60	**1.136**	**0.899**	5.079	3.929	1.187	0.928	8.054	7.120	1.212	0.955
	90	**1.224**	**0.968**	7.056	5.532	1.263	0.997	10.459	9.136	1.273	1.005
	180	**1.244**	**0.999**	10.512	8.572	1.279	1.004	13.720	12.019	1.321	1.053
	270	**1.247**	**0.979**	11.587	9.642	1.262	0.990	9.494	8.299	1.248	0.983
	360	**1.162**	**0.910**	10.951	8.945	1.345	1.069	×	×	1.258	0.997
	360[a]	**1.184**	**0.932**	8.398	7.440	1.419	1.138	10.320	8.885	1.240	0.974
	360[b]	**1.165**	**0.921**	10.265	8.574	1.357	1.069	1.463	1.163	1.198	0.941
Yellow Sea, China	30	**0.889**	**0.672**	2.443	1.914	0.929	0.712	3.699	3.290	1.039	0.799
	60	**1.018**	**0.785**	4.240	3.221	1.027	0.794	6.644	5.966	1.079	0.832
	90	**1.031**	**0.772**	5.921	4.321	1.060	0.816	8.633	7.640	1.130	0.866
	180	**1.187**	**0.947**	8.728	6.951	1.269	0.969	11.622	10.182	1.252	0.994
	270	**1.112**	**0.866**	8.480	6.745	1.201	0.925	9.028	8.055	1.137	0.885
	360	**1.026**	**0.883**	8.619	6.967	1.263	0.989	×	×	1.148	0.890
	360[a]	**1.063**	**0.826**	6.661	5.861	1.279	0.995	9.721	8.681	1.159	0.907
	360[b]	**1.049**	**0.807**	8.844	7.245	1.273	0.972	1.296	0.997	1.171	0.919
East China Sea	30	**0.847**	**0.644**	2.253	1.765	0.848	0.655	2.811	2.425	0.889	0.688
	60	**0.903**	**0.691**	3.553	2.653	0.953	0.738	5.017	4.479	0.918	0.716
	90	**0.923**	**0.713**	5.156	3.953	0.952	0.739	6.839	6.095	0.961	0.751
	180	**0.946**	**0.741**	6.702	5.571	0.971	0.758	9.741	8.692	0.992	0.772
	270	**0.944**	**0.738**	7.535	6.324	0.959	0.754	6.866	6.178	1.000	0.780
	360	**0.972**	**0.758**	7.061	5.750	1.000	0.783	×	×	0.980	0.771
	360[a]	**0.975**	**0.753**	5.075	4.468	0.997	0.784	8.267	7.303	0.994	0.779
	360[b]	**0.963**	**0.746**	6.598	5.423	1.087	0.862	1.139	0.891	0.975	0.760
Taiwan Strait	30	**1.112**	**0.850**	2.048	1.579	1.201	0.919	2.452	2.031	1.122	0.863
	60	**1.168**	**0.901**	2.587	2.126	1.193	0.919	3.767	3.210	1.176	0.915
	90	**1.191**	**0.912**	2.971	2.065	1.202	0.930	5.290	4.572	1.192	0.919
	180	**1.166**	**0.895**	4.869	3.786	1.169	0.901	6.975	6.073	1.283	0.987
	270	**1.175**	**0.906**	4.625	3.691	1.183	0.911	4.867	4.217	1.237	0.946
	360	**1.162**	**0.900**	5.272	4.205	1.203	0.925	×	×	1.186	0.910
	360[a]	**1.248**	**0.940**	3.674	3.121	1.253	0.974	6.695	5.869	1.240	0.948
	360[b]	**1.190**	**0.915**	4.671	3.764	1.296	1.004	1.386	1.043	1.283	0.988
South China Sea	30	**0.759**	**0.597**	1.367	1.110	0.767	0.599	1.599	1.305	0.832	0.669
	60	**0.809**	**0.651**	1.913	1.446	0.826	0.658	2.465	2.049	0.840	0.676
	90	**0.829**	**0.669**	2.367	1.803	0.850	0.676	3.044	2.571	0.875	0.702
	180	**0.861**	**0.691**	3.323	2.717	0.877	0.695	3.510	3.060	0.915	0.735
	270	**0.867**	**0.698**	3.643	3.037	0.897	0.717	3.168	2.706	0.887	0.714
	360	**0.877**	**0.702**	3.012	2.447	0.927	0.738	×	×	0.894	0.720
	360[a]	**0.847**	**0.680**	2.435	2.152	0.955	0.764	2.781	2.473	0.918	0.738
	360[b]	**0.891**	**0.705**	3.443	2.777	0.941	0.750	0.968	0.763	0.930	0.741
Count		**40**	**40**	0	0	0	0	0	0	0	0

Bold values highlights the optimal prediction results for each lead time in every sea area.

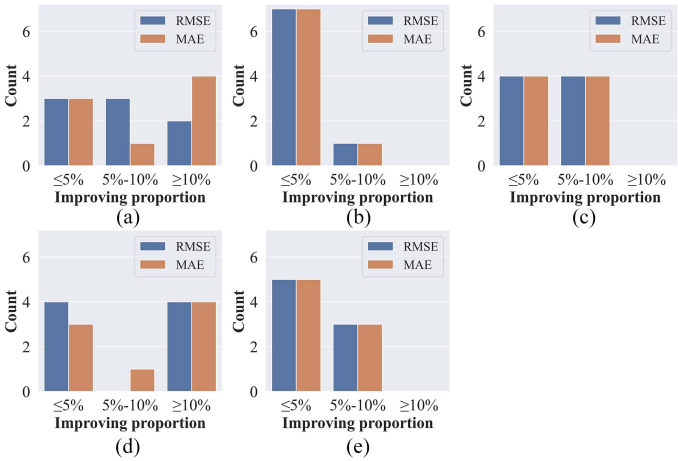

Fig. 8. **Univariate**, every ocean, all lead times, number of prediction improvement rate intervals of TransDtSt-Part relative to N-BEATS. (a) Yellow Sea. (b) East China Sea. (c) South China Sea. (d) Bohai Sea. (e) Taiwan Strait.

relatively small (RMSE∼0.75 °C–1 °C, MAE∼0.6 °C–0.8 °C), while those in the Bohai Sea and the Taiwan Strait are relatively large (RMSE∼0.95 °C–1.2 °C, MAE∼0.75 °C–0.9 °C). The prediction effect of the open sea is better than that of other oceans, which may be attributed to the larger SST fluctuations in the coastal area, while the SST in the open ocean area is relatively stable [17], [21], [22], [24].

Divided by sea area, the number of prediction improvement rates of TransDtSt-Part relative to N-BEATS is calculated for all lead times, as shown in Fig. 8. It can be found that the numbers of the three improvement rate intervals in the Yellow Sea and the Bohai Sea are roughly the same, while the improvement rates in the East China Sea, South China Sea, and Taiwan Strait are concentrated below 10%.

4) The prediction accuracy of the model TransDtSt-Part is significantly higher than that of the models LSTM for each ocean and each lead time. This may be due to the fact that LSTM adopts autoregressive decoding, so the model prediction error accumulates as the lead time increases. Since the period of the SST signal is about 360 days, the TCN prediction error is smaller than other lead times when seq_{len} is an integer multiple of 360 days (i.e., $\text{seq}_{\text{len}} = 720$ days). The predictive skill of TransDtSt-Part that considers full attention is better than that of Informer only with part of the attention coefficients.

2) Analysis of SST Multivariate Prediction Results: Similar to the analysis of univariate prediction results, the following conclusions on SST multivariate prediction results (Table II) are drawn.

1) The model TransDtSt-Part also has a trend of a steady and slow rise in performance as the lead time becomes longer.
2) In all research areas, the prediction error of TransDtSt-Part is smaller than that of other baseline models at all prediction horizons.

From Fig. 9, compared with N-BEATS, TransDtSt-Part improves more for longer lead times.

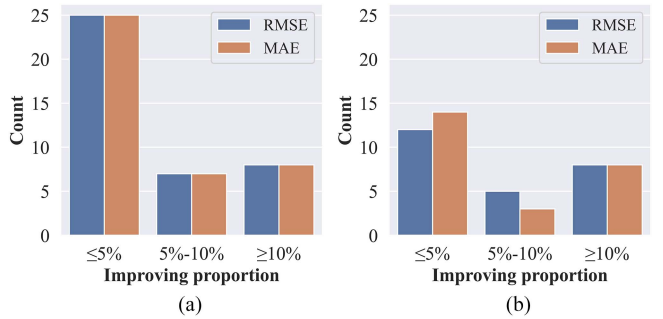

Fig. 9. **Multivariate**, all oceans, number of prediction improvement rate intervals of TransDtSt-Part relative to N-BEATS. (a) All lead times. (b) Lead times ≥ 180 days.

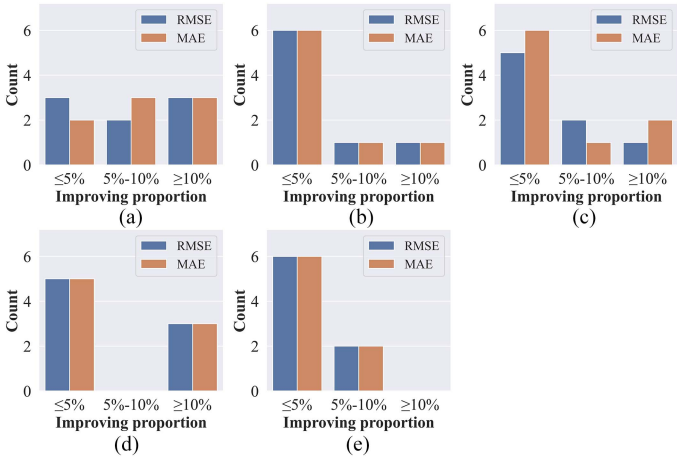

Fig. 10. **Multivariate**, every ocean, all lead times, number of prediction improvement rate intervals of TransDtSt-Part relative to N-BEATS. (a) Yellow Sea. (b) East China Sea. (c) South China Sea. (d) Bohai Sea. (e) Taiwan Strait.

3) The prediction errors in the open seas of the Yellow Sea, the East China Sea, and the South China Sea are relatively small (RMSE∼0.75 °C–1.20 °C, MAE∼0.60 °C–0.95 °C), while those in the Bohai Sea and the Taiwan Strait are relatively large (RMSE∼1 °C—1.25 °C, MAE∼0.80 °C–1.0 °C).

Fig. 10 exhibits the statistical value of the number of prediction improvement rate intervals of TransDtSt-Part relative to N-BEATS for all prediction horizons and research areas. The numbers of the three improvement rate intervals in the Yellow Sea are roughly equal, the improvement rates in the Bohai Sea are distributed at both ends of ≤5% and ≥10%, and the improvement rates in the East China Sea, South China Sea, and Taiwan Strait are concentrated below 10%.

3) Relationship Between Encoder Input Length and Model Performance: From Tables I and II, the prediction results of TransDtSt-Part present area-specific for $seq_{len} \in \{360, 540, 720\}$ days, $\text{label}_{\text{len}} = 360$ days, and $\text{pred}_{\text{len}} = 360$ days. Specifically, for the univariate case, the prediction errors of the Bohai Sea and the Yellow Sea decrease with the increase of seq_{len}, while the prediction errors of the East China Sea, the Taiwan Strait, and the South China Sea increase. For the multivariate case, the prediction error of the East China Sea

TABLE III
SEASONAL PREDICTION ERROR—**UNIVARIATE** PREDICTION SHOWCASE OF BOHAI SEA

Model	RMSE (°C)				MAE (°C)			
	Winter	Spring	Summer	Autumn	Winter	Spring	Summer	Autumn
TransDtSt-Part	0.857	0.823	0.816	0.865	0.711	0.673	0.658	0.705
LSTM	4.319	3.257	3.261	3.206	3.857	2.762	2.678	2.706
N-BEATS	0.902	0.872	0.826	0.897	0.735	0.721	0.666	0.722
TCN	3.195	2.845	3.454	3.048	2.735	2.431	3.109	2.556
Informer	0.901	0.804	0.843	0.925	0.743	0.654	0.676	0.747

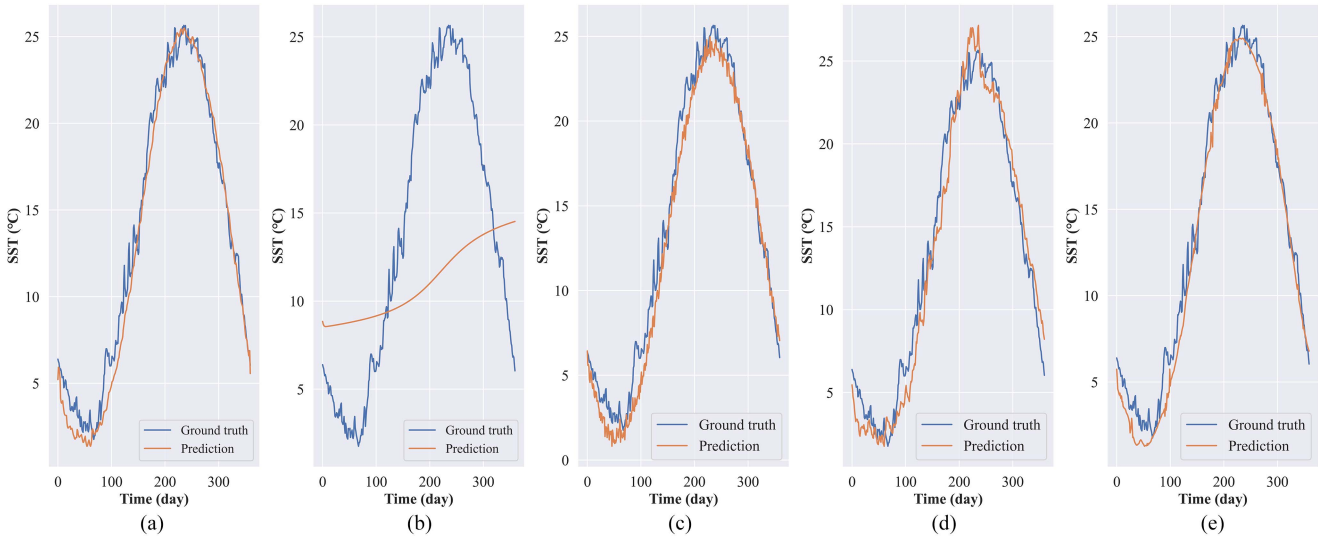

Fig. 11. Under the **univariate** pattern, the predictions ($\text{pred}_{\text{len}} = 360$ days with $\text{seq}_{\text{len}} = 720$ days) of (a) TransDtSt-Part, (b) LSTM, (c) N-BEATS, (d) TCN, and (e) informer on the Bohai Sea. The orange/blue curves stand for slices of the prediction/ground truth.

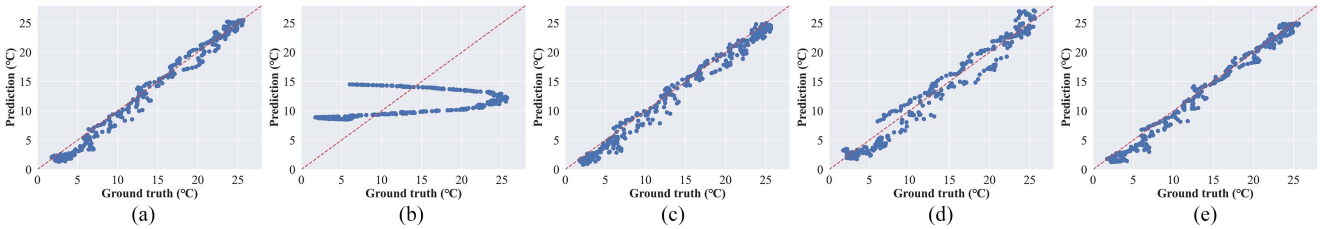

Fig. 12. Under the **univariate** pattern, the correlations between the predictions ($\text{pred}_{\text{len}} = 360$ days with $\text{seq}_{\text{len}} = 720$ days) of (a) TransDtSt-Part, (b) LSTM, (c) N-BEATS, (d) TCN, and (e) informer and the ground truth on the Bohai Sea. The dark red dash line stands for the 1:1 line.

decreases with the seq_{len} increasing. The prediction errors of the Bohai Sea, the Yellow Sea, and the Taiwan Strait increase first and then decrease. But the South China Sea is the opposite, first decreasing and then increasing.

4) Univariate Prediction Showcase: Fig. 11 shows the prediction slices of the model TransDtSt-Part and the baseline models in the Bohai Sea with $\text{seq}_{\text{len}} = 720$ days, $\text{label}_{\text{len}} = 360$ days, and $\text{pred}_{\text{len}} = 360$ days. Fig. 12 shows the correlation between each model's prediction and ground truth.

It can be seen from Fig. 11(b) that due to the autoregressive decoding, the LSTM prediction error continues to accumulate, and the prediction curve seriously deviates from the ground truth. From Figs. 11(c), (d), and 12(c), (d), N-BEATS and TCN have larger prediction errors at some troughs (2 °C–5 °C) and peaks (22 °C–25 °C). Although in Figs. 11(a), (e), and 12(a), (e), informer and TransDtSt-Part have relatively poor performance in the low-value part of SST, they will be more in line with the ground truth in the high-value part of SST. Compared with informer, the overall predictive skill of TransDtSt-Part considering all attention coefficients is better.

5) Multivariate Prediction Showcase: Taking the South China Sea as an example, when the pred_{len} is 270 days, Fig. 13 exhibits the slices of the last dimension of the prediction results

TABLE IV
SEASONAL PREDICTION ERROR—**MULTIVARIATE** PREDICTION SHOWCASE OF SOUTH CHINA SEA

Model	RMSE (°C)			MAE (°C)		
	Winter	Spring	Summer	Winter	Spring	Summer
TransDtSt-Part	1.069	1.045	1.072	1.084	0.842	0.833
LSTM	11.938	10.977	12.439	11.009	9.791	9.413
N-BEATS	1.357	1.344	1.342	1.380	1.091	1.073
TCN	1.482	1.471	1.452	1.446	1.206	1.198
Informer	1.220	1.139	1.155	1.155	0.974	0.914

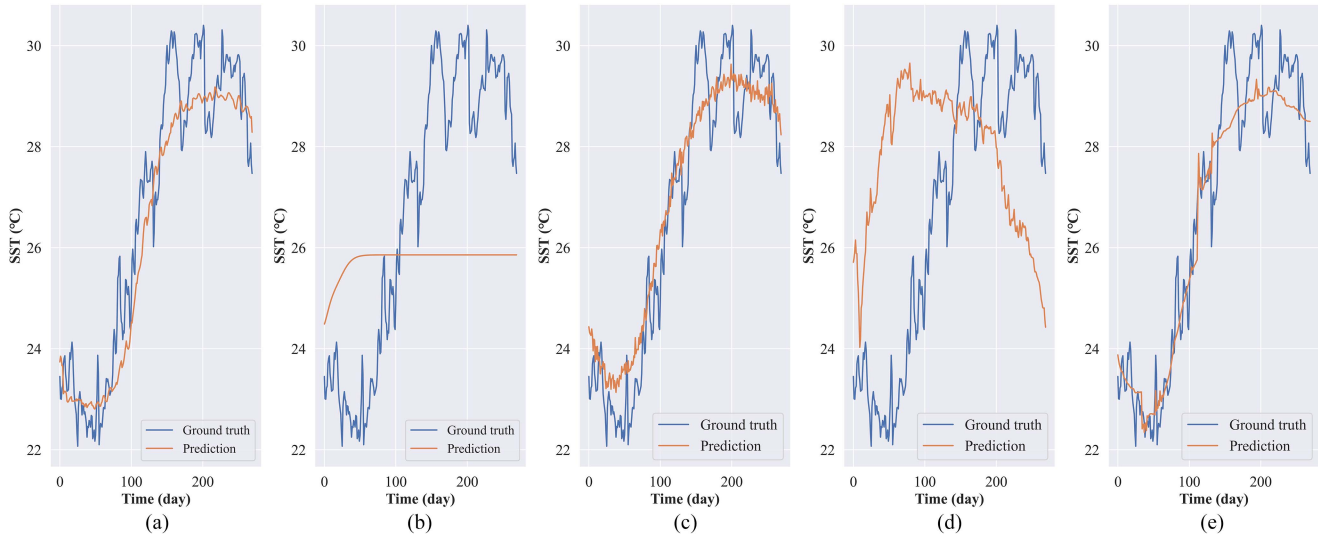

Fig. 13. Based on the **multivariate** pattern, the predictions ($\text{pred}_{\text{len}} = 270$ days with $\text{seq}_{\text{len}} = 360$ days) of (a) TransDtSt-Part, (b) LSTM, (c) N-BEATS, (d) TCN, and (e) informer on the South China Sea. The orange/blue curves stand for slices of the prediction/ground truth.

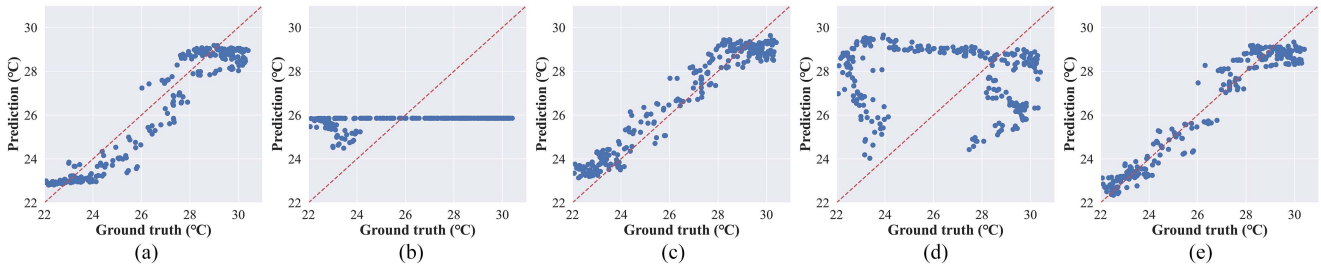

Fig. 14. Based on the **multivariate** pattern, the correlations between the predictions ($\text{pred}_{\text{len}} = 270$ days with $\text{seq}_{\text{len}} = 360$ days) of (a) TransDtSt-Part, (b) LSTM, (c) N-BEATS, (d) TCN, and (e) informer and the ground truth on the South China Sea. The dark red dash line stands for the 1:1 line.

and the corresponding Ground truth. Fig. 14 shows the correlation between the prediction and Ground truth at this time.

The prediction performance of LSTM in Figs. 13(b) and 14(b) is even worse, and the model cannot make normal predictions and finally presents a straight line. TCN has also been unable to capture the individual long-range dependencies between outputs and inputs for SST long-sequence [Figs. 13(d) and 14(d)]. From Figs. 13(c) and 14(c), N-BEATS exhibits overestimation at the lower SST (22 °C–24 °C) and underestimation at the higher SST (28 °C–30.5 °C) under this case. In Fig. 13(a) and (e), informer and TransDtSt-Part can still accurately grasp the long-term change trend of SST. On the whole, informer's prediction is more fluctuating, and the prediction curve of TransDtSt-Part is smoother. Both informer and TransDtSt-Part have large prediction errors in the range of 24 °C–28 °C [Fig. 14(a) and (e)].

6) Seasonal Prediction Error Analysis: For all SST prediction slice data in the univariate prediction showcase (i.e., Bohai Sea with $\text{seq}_{\text{len}} = 720$ days, $\text{label}_{\text{len}} = 360$ days, and $\text{pred}_{\text{len}} = 360$ days) and multivariate prediction showcase (i.e., South China Sea with $\text{seq}_{\text{len}} = 360$ days, $\text{label}_{\text{len}} = 360$ days, and

$\text{pred}_{\text{len}} = 270$ days), according to the four seasons of Winter (Jan.–Mar.), Spring (Apr.–Jun.), Summer (Jul.–Sep.), and Autumn (Oct.–Dec.), the average predictive skills of TransDtSt-Part and the baseline models are calculated, respectively, as shown in Tables III and IV.

It is easy to find from Tables III and IV that except for the Spring in the univariate prediction of Bohai Sea, the prediction performance of the TransDtSt-Part model is slightly inferior to that of the informer and its predictive skills are the best in other seasons. It proves the excellent seasonal SST prediction ability of the TransDtSt-Part model.

V. Conclusion

We focus on the long-term prediction of SST in the China Sea at a fine-grained daily level in the article. Transformer has powerful time-series modeling capabilities, but it also with some disadvantages, such as high computational complexity, low autoregressive decoding efficiency, and easy accumulation of errors. We make targeted improvements to build the model TransDtSt-Part: using generative decoding, embedding time-dimensional information, and introducing attention distilling and partial stacked connection. Among the extensive experiments of two prediction patterns and multiple lead times in the five China Sea regions, the prediction performance of the model TransDtSt-Part outperforms all competitive baseline models to varying degrees, proving its excellent long-term predictive skill of SST. It may be helpful for many urgent long-term requirements in marine and climate applications.

Acknowledgment

Special thanks for the support from high-resolution SST data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA. All Daily mean SST data used in this article were obtained from NOAA/OAR/ESRL PSL at https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html.

We would also like to thank the third-party Python library Darts and Optuna.

References

[1] M. Bouali, O. T. Sato, and P. S. Polito, "Temporal trends in sea surface temperature gradients in the South Atlantic Ocean," *Remote Sens. Environ.*, vol. 194, pp. 100–114, Jun. 2017, doi: 10.1016/j.rse.2017.03.008.

[2] T. D. Herbert, L. C. Peterson, K. T. Lawrence, and Z. H. Liu, "Tropical ocean temperatures over the past 3.5 million years," *Science*, vol. 328, no. 5985, pp. 1530–1534, Jun. 2010, doi: 10.1126/science.1185435.

[3] K. Patil, M. C. Deo, S. Ghosh, and M. Ravichandran, "Predicting sea surface temperatures in the North Indian Ocean with nonlinear autoregressive neural networks," *Int. J. Oceanogr.*, vol. 2013, Apr. 2013, Art. no. 302479, doi: 10.1155/2013/302479.

[4] K. Patil, M. C. Deo, and M. Ravichandran, "Prediction of sea surface temperature by combining numerical and neural techniques," *J. Atmos. Ocean. Technol.*, vol. 33, no. 8, pp. 1715–1726, Aug. 2016, doi: 10.1175/JTECH-D-15-0213.1.

[5] T. N. Krishnamurti, A. Chakraborty, R. Krishnamurti, W. K. Dewar, and C. A. Clayson, "Seasonal prediction of sea surface temperature anomalies using a suite of 13 coupled atmosphere-ocean models," *J. Climate*, vol. 19, pp. 6069–6088, Dec. 2006, doi: 10.1175/JCLI3938.1.

[6] T. N. Stockdale, M. A. Balmaseda, and A. Vidard, "Tropical Atlantic SST prediction with coupled ocean–atmosphere GCMs," *J. Climate*, vol. 19, no. 23, pp. 6047–6061, Dec. 2006, doi: 10.1175/JCLI3947.1.

[7] Y. F. Wang, Z. H. Zhang, and P. Huang, "An improved model-based analogue forecasting for the prediction of the tropical Indo-Pacific sea surface temperature in a coupled climate model," *Int. J. Climatol.*, vol. 40, no. 15, pp. 6346–6360, Dec. 2020, doi: 10.1002/joc.6584.

[8] J. S. Kug, I. S. Kang, J. Y. Lee, and J. G. Jhun, "A statistical approach to Indian Ocean sea surface temperature prediction using a dynamical ENSO prediction," *Geophys. Res. Lett.*, vol. 31, no. 9, May 2004, Art. no. L09212, doi: 10.1029/2003GL019209.

[9] R. S. Neetu, S. Basu, A. Sarkar, and P. K. Pal, "Data-adaptive prediction of sea-surface temperature in the Arabian Sea," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 9–13, Jan. 2011, doi: 10.1109/LGRS.2010.2050674.

[10] S. G. Aparna, G. D'Souza, and N. B. Arjun, "Prediction of daily sea surface temperature using artificial neural networks," *Int. J. Remote Sens.*, vol. 39, no. 12, pp. 4214–4231, Apr. 2018, doi: 10.1080/01431161.2018.1454623.

[11] S. Hou et al., "D2CL: A dense dilated convolutional LSTM model for sea surface temperature prediction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12514–12523, 2021, doi: 10.1109/JSTARS.2021.3128577.

[12] S. Y. Hou et al., "MUST: A multi-source spatio-temporal data fusion model for short-term sea surface temperature prediction," *Ocean Eng.*, vol. 259, Sep. 2022, Art. no. 111932, doi: 10.1016/j.oceaneng.2022.111932.

[13] M. Jahanbakht, W. Xiang, and M. R. Azghadi, "Sea surface temperature forecasting with ensemble of stacked deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 1502605, doi: 10.1109/LGRS.2021.3098425.

[14] J. J. Liu, B. G. Jin, J. K. Yang, and L. Y. Xu, "Sea surface temperature prediction using a cubic B-spline interpolation and spatiotemporal attention mechanism," *Remote Sens. Lett.*, vol. 12, no. 5, pp. 478–487, May 2021, doi: 10.1080/2150704X.2021.1897182.

[15] K. Patil and M. C. Deo, "Prediction of daily sea surface temperature using efficient neural networks," *Ocean Dyn.*, vol. 67, pp. 357–368, Apr. 2017, doi: 10.1007/s10236-017-1032-9.

[16] K. Patil and M. C. Deo, "Basin-scale prediction of sea surface temperature with artificial neural networks," *J. Atmos. Ocean. Technol.*, vol. 35, no. 7, pp. 1441–1455, Jul. 2018, doi: 10.1175/JTECH-D-17-0217.1.

[17] L. Wei, L. Guan, and L. Q. Qu, "Prediction of sea surface temperature in the South China Sea by artificial neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 558–562, Apr. 2020, doi: 10.1109/LGRS.2019.2926992.

[18] S. Wolff, F. O'Donnch, and B. Chen, "Statistical and machine learning ensemble modelling to forecast sea surface temperature," *J. Mar. Syst.*, vol. 208, Aug. 2020, Art. no. 103347, doi: 10.1016/j.jmarsys.2020.103347.

[19] C. J. Xiao et al., "A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data," *Environ. Model. Softw.*, vol. 120, Oct. 2019, Art. no. 104502, doi: 10.1016/j.envsoft.2019.104502.

[20] 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, doi: 10.1109/LGRS.2019.2931728.

[21] L. Y. Xu, Q. Li, J. Yu, L. Wang, J. Xie, and S. X. Shi, "Spatiotemporal predictions of SST time series in China's offshore waters using a regional convolution long short-term memory (RC-LSTM) network," *Int. J. Remote Sens.*, vol. 41, no. 9, pp. 3368–3389, May 2020, doi: 10.1080/01431161.2019.1701724.

[22] L. Y. Xu, Y. F. Li, J. Yu, Q. Li, and S. X. Shi, "Prediction of sea surface temperature using a multiscale deep combination neural network," *Remote Sens. Lett.*, vol. 11, no. 7, pp. 611–619, Jul. 2020, doi: 10.1080/2150704X.2020.1746853.

[23] Y. Yang, J. Dong, X. Sun, E. Lima, Q. Mu, and X. H. Wang, "A CFCC-LSTM model for sea surface temperature prediction," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 207–211, Feb. 2018, doi: 10.1109/LGRS.2017.2780843.

[24] Q. Zhang, H. Wang, J. Dong, G. Zhong, and X. Sun, "Prediction of sea surface temperature using long short-term memory," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1745–1749, Oct. 2017, doi: 10.1109/LGRS.2017.2733548.

[25] X. Zhang, Y. Li, A. C. Frery, and P. Ren, "Sea surface temperature prediction with memory graph convolutional networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2022, Art. no. 8017105, doi: 10.1109/LGRS.2021.3097329.

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[27] C. A. Huang et al., "Music transformer: Generating music with long-term structure," in *Proc. Int. Conf. Learn. Representations*, 2019, doi: 10.48550/arXiv.1809.04281.
[28] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010, doi: 10.48550/arXiv.1706.03762.
[29] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. ICLR*, 2016, doi: 10.48550/arXiv.1511.07289.
[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[31] S. J. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, doi: 10.48550/arXiv.1803.01271.
[32] H. Y. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI*, Virtual Event, 2021, pp. 11106–11115, doi: 10.48550/arXiv.2012.07436.
[33] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *Proc. ICLR*, 2020, doi: 10.48550/arXiv.1905.10437.

Hao Dai was born in 1982. He received the Ph.D. degree in instrument science and technology from the Zhejiang University, Hangzhou, China, in 2012.

He is currently a Senior Engineer with the Institute of Ocean Exploration Technology, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China. His research interests include artificial intelligence oceanography and ocean metrology.

Zhigang He was born in 1973. He received the Ph.D. degree in ocean chemistry from the Xiamen University, Xiamen, China, in 2005.

He is currently an Associate Professor with the Xiamen University. His research interests include ocean observation technology, circulation in the South China Sea and its adjacent region.

Guomei Wei was born in 1985. She received the M.S. degree in environmental sciences from the Xiamen University, Xiamen, China, in 2010.

Her research interests include quality control technology for radar data.

Famei Lei was born in 1984. He received the M.S. degree in environmental sciences from the Xiamen University, Xiamen, China, in 2012.

His research interest includes data processing of ocean observation instruments.

Xining Zhang was born in 1982. She received the Ph.D. degree in optical engineering from the Zhejiang University, Hangzhou, China, in 2012.

She is currently an Associate Professor with the College of Information Science and Engineering, Huaqiao University, Quanzhou, China. Her research interests include surface plasmon properties in metal micro/nano structures, and optical devices based on micro/nanofibers.

Weijie Zhang was born in 1986. She received the M.S. degree in environmental sciences from the Xiamen University, Xiamen, China, in 2015.

Her research interest includes visualization in ocean science.

Shaoping Shang was born in 1962. He received the M.S. degree in physical oceanography from the Xiamen University, Xiamen, China, in 1986.

He is currently a Professor with the College of Ocean and Earth Sciences, Xiamen University, Xiamen, China. His research interests include ocean observation, marine information technology, numerical modeling, and remote sensing of marine environment.