Abstract: In this study, we investigate the feasibility of using historical remote sensing data to predict the future three-dimensional subsurface ocean temperature structure. We also compare the performance differences between predictive models and real-time reconstruction models. Specifically, we propose a multi-scale residual spatiotemporal window ocean (MSWO) model based on a spatiotemporal attention mechanism, to predict changes in the subsurface ocean temperature structure over the next six months using satellite remote sensing data from the past 24 months. Our results indicate that predictions made using historical remote sensing data closely approximate those made using historical in situ data. This finding suggests that satellite remote sensing data can be used to predict future ocean structures without relying on valuable in situ measurements. Compared to future predictive models, real-time three-dimensional structure reconstruction models can learn more accurate inversion features from real-time satellite remote sensing data. This work provides a new perspective for the application of artificial intelligence in oceanography for ocean structure reconstruction.

Keywords: temperature structure prediction; temperature structure reconstruction; spatiotemporal window ocean; satellite observations; spatiotemporal attention mechanism

1. Introduction

The ocean plays a critical role in regulating the stability of the Earth’s system by absorbing a significant portion of global heat. Temperature, as one of the most fundamental marine physical quantities, is intricately linked to the density structure of the ocean [1,2]. This relationship influences not only the flow field but also biological activities and chemical reactions within the marine environment [3]. Recent studies have demonstrated a notable upward trend in the heat content of the upper ocean over the past few decades [4]; such increases in ocean temperature are associated with the potential for meteorological disasters, including typhoons and storm surges [5]. Consequently, investigating and understanding changes in ocean temperature structure are essential for promoting marine environmental awareness, ecological protection, and disaster prevention.

Numerical simulation is a traditional method employed to obtain the three-dimensional structure of the ocean and predict its dynamic processes. Various models exhibit distinct
advantages, for instance, the Hybrid Coordinate Ocean Model (HYCOM) [6] features a significant vertical hierarchical structure, establishing a practical hybrid vertical coordinate system. The Finite-Volume Coastal Model (FVCOM) [7] accurately fits the coastline boundary and seabed topography using specific conservation equations. Despite their utility, these models rely on dynamic simplification processes and settings of various model parameters, such as bottom friction coefficient and eddy viscosity coefficients. They thus may not capture the full complexity and variability of the real ocean. As a result, the accuracy of numerical simulation data can be limited [8]. Additionally, the high computational overhead associated with these simulations poses a significant challenge. To address these limitations, numerous research institutions employ various instruments such as subsurface mooring, large buoys, drifting buoys, and gliders for real-time observation of the three-dimensional ocean structure [9–11]. However, field measurement methods face inherent challenges, including high sampling costs, difficulty in data acquisition, and low spatiotemporal resolution [12,13]. Satellite remote sensing offers high resolution, large-scale coverage, and long-term continuity but is restricted to detecting surface information due to transmission media limitations [14,15]. To overcome this, researchers have proposed data-driven deep learning strategies to reconstruct subsurface ocean structures from satellite data, a process termed deep ocean remote sensing (DORS) [16,17]. Recent years have seen DORS emerge as an efficient and innovative approach to obtaining the three-dimensional structure of the ocean.

There are currently four major schemes for DORS work (Figure 1). Figure 1a shows the use of known satellite remote sensing data of the ocean surface to predict future changes at the ocean surface. For instance, Zhang et al. [18] employed the Long Short-Term Memory (LSTM) method to forecast sea surface temperature in offshore China, He et al. [19] developed the DSL method for predicting sea surface temperature, Zhang et al. [20] devised a U-Net method for time series prediction of sea surface salinity in the Western Pacific Ocean, Xu et al. [21] combined Memory in Memory (MIM) and Variational Mode Decomposition (VMD) to propose the VMD-MIM model, which further enhances the prediction performance of sea surface temperature. These approaches primarily focus on predicting surface changes, whereas subsurface data are often more valuable to researchers [22].

Figure 1b depicts the use of historical subsurface three-dimensional structure data to forecast future subsurface changes. For example, Sun et al. [23] proposed a 3D U-Net model, utilizing the past 12 months of subsurface data to predict the subsurface structure for the next 12 months. Yue et al. [24] introduced the SA-PredRNN model, based on SA-ConvLSTM and PredRNN, to achieve similar predictions using historical subsurface data. Although these predictive models are valuable, they rely on extensive measured data as input, which poses practical limitations due to the data’s scarcity and acquisition challenges.

Figure 1c represents the most extensively studied approach, focusing on the reconstruction of subsurface ocean structures using remote sensing data. Su et al. [25] explored the reconstruction of subsurface temperature and salinity anomalies using machine learning methods such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost), while Meng et al. [26] achieved high-resolution subsurface reconstruction through Convolutional Neural Networks (CNNs). Xie et al. [27] implemented subsurface reconstruction in the South China Sea region using a U-Net model based on convolutional neural networks. Recently, the introduction of the Transformer framework and attention mechanisms has facilitated the efficient reconstruction of subsurface structures by learning from historical data. For instance, Zhang et al. [28] developed a discrete point historical time series model to invert subsurface structures by capturing temporal changes, while Su et al. [29] utilized the Transformer framework to sample and extract spatial features from images of different scales, reducing computational costs and enabling efficient reconstruction of subsurface density fields. These methods, however, are focused on reconstruction and do not predict future subsurface structures.

Figure 1d illustrates the relatively novel approach of using historical remote sensing data to predict future subsurface changes in the ocean. In 2024, Liu et al. [22] introduced
an artificial intelligence model that employed historical remote sensing images to forecast future subsurface ocean structures, training the model with satellite remote sensing and numerical simulation data, and achieving notable results in the South China Sea region.

By comparing these DORS methods, we identify several pertinent questions for further exploration: (1) Is there a performance difference between using historical remote sensing data to predict future subsurface ocean profile structures and using historical subsurface ocean profile structures for the same purpose? (2) What is the relationship between the reconstruction of subsurface structures that combines historical and real-time data, and the prediction of subsurface structures using only historical data?

In this paper, we propose the multiscale spatiotemporal window ocean (MSWO) model, which combines the features of the spatiotemporal series network SimVP [30] and the computer vision Swin Transformer [31] network. The MSWO model employs a window attention mechanism in the spatiotemporal dimension to achieve low computational-load attention effects. Additionally, it extracts the nonlinear spatiotemporal relationships in the data through multi-scale residuals. To better represent the spatial and temporal characteristics of the data, we introduce global location coding during information extraction and use the channel attention mechanism to extract key feature information from remote sensing images. The model was validated against measured Argo grid data in the Central South Pacific. Historical remote sensing data from the past 24 months were used to predict the subsurface temperature structure trends in the Central South Pacific for the next 6 months. Furthermore, we explored the effect of using measured profile structures from the past 24 months to predict the subsurface structure for the next 6 months, comparing the reconstructed results with the predicted results on the test dataset. The results indi-

Figure 1. (a) Surface ocean prediction using remote sensing images. (b) Subsurface ocean prediction using measured profile data. (c) Subsurface ocean reconstruction using satellite remote sensing images. (d) Subsurface ocean prediction using satellite remote sensing images.
cate that our model achieves the highest inversion accuracy in both reconstruction and prediction processes. This method presents a promising new approach for transparent ocean observation.

2. Related Work

Both the reconstruction of ocean subsurface structures and the prediction of future subsurface changes can be accomplished using spatiotemporal series models. With the development of artificial intelligence, more and more remote sensing images are used to reconstruct ocean subsurface structures and begin to invert historical data as one of the input factors [28,29].

Spatiotemporal series models have their origins in single-element time series prediction models, with the most traditional method being the recurrent neural networks (RNNs). The LSTM [32] subsequently improved the model’s efficiency in learning temporal patterns. In 2015, Shi et al. [33] advanced this field by introducing the ConvLSTM model, which replaced the fully connected layer with a convolutional layer in the FC-LSTM, marking a transition from time series models to spatiotemporal series models. This innovation extended LSTM input data to multidimensional images, maintaining the fundamental principles of LSTM and other recurrent neural networks. The prediction for the next time point is achieved by iterating through the forgetting gate, output gate, and memory cell. In 2017, Wang et al. [34] proposed PredRNN, based on ConvLSTM, which introduced a long-term memory module to enhance the accuracy of spatiotemporal series prediction tasks.

The introduction of the attention mechanism and the Transformer framework [35] in 2017 revolutionized natural language-processing research. Researchers discovered that the attention mechanism’s ability to capture global changes was superior to that of convolutional neural networks, leading to its application in computer vision. Various Transformer framework variants [31,36–38] have since achieved remarkable success in object detection, image classification, and semantic segmentation. The attention mechanism’s capability to capture global information has also been applied to time series prediction tasks [39,40]. In 2020, Lin et al. [41] incorporated the self-attention module into the ConvLSTM model, resulting in the SA-ConvLSTM model, which can capture global spatiotemporal correlations. However, with higher-resolution image inputs, spatiotemporal series prediction models faced similar computational cost challenges as those in computer vision. The Swin Transformer [31] improved model performance while significantly reducing computational complexity. Consequently, in 2023, Tang et al. [42] replaced the convolutional layer in ConvLSTM with the Swin Transformer, achieving optimal results on video prediction datasets such as Moving MINST.

However, the iterative RNN-based spatiotemporal series models have inherent flaws. As illustrated in Figure 2a, the prediction for the current time step in an iterative model is influenced by the previous time step’s prediction, leading to error accumulation over time. Additionally, this process tends to make the model overly reliant on immediate past context information, hindering long-term information learning [43]. To address this issue, various generative models have been developed to generate all prediction results by capturing global information in a single step [39,44], as shown in Figure 2b. In 2023, Tan et al. [30] proposed the SimVP model, which uses an Encoder–Decoder architecture for feature extraction and large-kernel convolution to simulate a global attention mechanism, achieving excellent results in spatiotemporal prediction. With the advent of a new generation of satellites, remote sensing images now feature larger scales and higher resolutions. Consequently, the primary objective of this paper is to design an efficient artificial intelligence model that supports high-resolution images to predict the future three-dimensional structure of the ocean.
Figure 2. (a) Flow chart of RNN iterative framework on spatiotemporal prediction tasks. (b) Flow chart of the generative framework for spatiotemporal prediction tasks.

3. Method

3.1. Overall Architecture

The MSWO model aims to design an autoencoder architecture suitable for spatiotemporal sequence prediction tasks, and predicts future images based on the learning of past temporal related frames. Different from the traditional iterative RNN, the MSWO model uses a generative strategy. The traditional automatic encoder generates a single step frame in static time and minimizes the difference between the true probability distribution $P_s$ and the predicted probability distribution $P'_s = G_{θ}(X)$ by mapping $G_{θ} : X \rightarrow \hat{Y}$. The optimal parametric neural network is as follows:

$$\theta^* = \arg \min_{θ} \text{Div}(P_s, P'_s), \quad (1)$$

where $\text{Div}$ represents a specific loss function. Similarly, MSWO encodes the frames of the entire historical time dimension, decodes the future prediction time dimension, and minimizes the difference between the predicted and true probability distributions by mapping $F_{θ} : X^{t,T} \rightarrow \hat{Y}^{t+1,T'}$. The optimal parametric neural network is:

$$\theta^* = \arg \min_{θ} \sum_{t=1}^{t+1+T'} \text{Div}(P'_s, P'_s), \quad (2)$$

In the experiment, we choose the $L_2$ loss function to minimize the error between this mapping:

$$\theta^* = \arg \min_{θ} \sum_{t=1}^{t+1+T'} \| Y'_t - F_{θ}(X^{t}_t) \|^2 \quad (3)$$

3.2. MSWO Module

Assuming that the size of the input remote sensing image is $H \times W \times C$ in $T$ times, $H$ represents the length of the image, $W$ represents the width of the image, and $C$ represents the number of channels of the image, the entire input can be represented by a tensor $\mathcal{X} \in \mathbb{R}^{T \times H \times W \times C}$. As shown in Figure 3, the input shape is $B \times T \times C \times H \times W$, where $B$ represents batch size. The whole MSWO model consists of three parts: spatial encoder, window attention, and spatial decoder.
The spatial encoder consists of two subsampled convolution layers, the convolutional block attention module (CBAM) [45], global attention coding [46], and residual concatenation. By stacking convolution layers, high-dimensional satellite remote sensing images are encoded into the low-dimensional potential space as follows:

$$z = \delta(\text{Norm(Conv}(z))),$$  

(4)

where $\delta$ represents the activation function used to extract nonlinear variations from features, and $\text{Norm}$ is the normalization layer. CBAM includes a channel attention module and a spatial attention module, which use average pooling and max pooling methods to extract useful channel information and spatial information, helping the model to better learn critical information and enhance sensitivity, thereby improving the overall performance of the model.

After spatial encoder, we reshape the dimensions of the tensor as $B \times (T \times C) \times H^* \times W^*$. The purpose of this is to stack the single frame images at different times along the time axis and update the channel of the image to the combination of the doped time dimension. The tensor is then entered into the window attention module. For large-scale and high-resolution spatiotemporal series prediction tasks, the computational complexity of global attention is generally $\Omega(MSA_1) = 4HWCT^2 + 2(HWC)^2T$, and the spatial complexity is $O((HWC)^2)$. Such square-level computational complexity causes a very large load on intensive prediction tasks [36]. By stacking time dimension and channel dimension, MSWO controls the matrix multiplication operation in smaller dimensions. The overall computational complexity of the model is $\Omega(MSA)_2 = 4HW(CT)^2 + 2(HW)^2CT$, and the space complexity becomes $O((HW)^2)$, which greatly reduces the overhead of the model. In addition, inspired by the Swin Transformer [35] and SwinLSTM [40] models, we designed SWO and MSWO options in the window attention section, as shown in Figure 4c,d. SWO includes general window segmentation and a sliding window module. By dividing the whole image and performing attention calculation in the window, the computing overhead is further reduced. At this time, the whole image is divided into
\( \frac{H}{M} \times \frac{W}{M} \) window matrices of \( M \times M \) size, and the computational complexity is reduced to \( \Omega(MSA)^2 = \frac{HW}{M^2} \times \left( 4M^2(CT)^2 + 2M^4 CT \right) = 4HW(CT)^2 + 2M^2 HWCT \), and the space complexity is \( O(M^4) \). In MSWO, we use patch merging to acquire multi-scale image data, and through the residual joining operation, we can focus on a larger sensory field in the \( M \times M \) size window.

![Diagram](image.png)

**Figure 4.** Details in the MSWO process. (a) The process of tensor dimension change in the whole process. (b) Window segmentation and attention calculation diagram. (c) SWO model flow chart. (d) MSWO model flow chart.

The Swin transformer cell is divided into two modules. The first module captures the token relationship inside the window through the window-multi-head attention mechanism (W-MHA); the second module is the same as the first module, only the W-MHA is replaced by the shift window-multi-head attention (SW-MHA) to capture the token relationship between adjacent windows. Finally, the decoding process also needs to reshape the tensor shape into \((B \times T) \times C \times H^* \times W^*\) to extract the information of a single image, and map the subsampled spatiotemporal data back to the original size through a deconvolution operation as follows:

\[
    z = \delta(Norm(ConvTrans(z)))
\]

(5)

**4. Experiments**

In the experimental part, we introduce the whole experimental design of the model, including data sources, processing methods, parameters of the training model, evaluation indexes of the model results, and the design of the ablation experiment.

**4.1. Data**

In this study, we utilized globally available datasets, including sea surface temperature (SST), absolute dynamic topography (ADT), sea surface salinity (SSS), and Argo-measured datasets. The temporal resolution for all data is monthly averages, and the spatial resolution was interpolated to \(1^\circ \times 1^\circ\) using interpolation methods. During the training and testing phases, satellite remote sensing data were integrated as input variables, and Argo profile data served as output variables. As detailed in Table 1, SST data were sourced from the National Oceanic and Atmospheric Administration (NOAA)’s OISST dataset, ADT data from Aviso, and SSS data from SMOS satellite’s Level 3 data. The study area encompassed the Central South Pacific, spanning from 44.5°S to 18.5°N and 116.5°W to 179.5°W. The input dimensions for the remote sensing images were 64 × 64 pixels. The temporal sequence...
of the dataset ranged from January 2011 to December 2019, encompassing 108 months. The time series data were divided into training, validation, and testing sets in a 7:1:1 ratio, employing a sliding window method [39]. The input time series covered 24 months, with predictions extending 6 months into the future. Argo depth selection focused on the upper 250 m of the ocean, divided into 10 depth layers (10, 20, 30, 50, 75, 100, 125, 150, 200, and 250 m). Outlier data were processed and optimally interpolated, and the following method was used for data normalization:

\[
X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}
\]  

The MSWO model was implemented using the PyTorch framework. The training process utilized the \(L_2\) loss function and the Adam optimizer with a learning rate of 0.001. The model was trained for 1000 epochs, with an early stopping criterion set at 50 epochs, and a batch size of 2. To ensure robustness of the final results, all experiments were repeated three times on an Nvidia Tesla V100 GPU, with the mean and standard deviation of these repetitions reported as the final results.

### Table 1. Data sources and resolutions used in present study.

<table>
<thead>
<tr>
<th>Data</th>
<th>Data Source</th>
<th>Spatial Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argo</td>
<td><a href="http://apdrc.soest.hawaii.edu/projects/Argo/">http://apdrc.soest.hawaii.edu/projects/Argo/</a> (accessed on 1 May 2024)</td>
<td>1°, monthly</td>
</tr>
<tr>
<td>ADT</td>
<td><a href="https://www.aviso.altimetry.fr/en/data/products/">https://www.aviso.altimetry.fr/en/data/products/</a> (accessed on 1 May 2024)</td>
<td>0.25°, monthly</td>
</tr>
<tr>
<td>SST</td>
<td><a href="https://climatedataguide.ucar.edu/climate-data/sst-data-noaa-high-resolution-025x025-blended-analysis-daily-sst-and-ice-oisstv2">https://climatedataguide.ucar.edu/climate-data/sst-data-noaa-high-resolution-025x025-blended-analysis-daily-sst-and-ice-oisstv2</a> (accessed on 1 May 2024)</td>
<td>0.25°, monthly</td>
</tr>
<tr>
<td>SSS</td>
<td><a href="https://www.catds.fr/Products/Catalogue-CPDC/Catds-products-from-Sextant#metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b">https://www.catds.fr/Products/Catalogue-CPDC/Catds-products-from-Sextant#metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b</a> (accessed on 1 May 2024)</td>
<td>25 km, monthly</td>
</tr>
</tbody>
</table>

### 4.2. Evaluation Indicator

We used three evaluation metrics to evaluate the performance of the predicted results against the measured results, which are mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE). These three indicators are used to measure the correlation between two data vectors, and the calculation process is as follows:

\[
MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2,
\]  

\[
RMSE = \sqrt{MSE},
\]  

\[
MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|,
\]

where, \(y_i\) and \(\hat{y}_i\) represent the true value and reconstructed value, respectively, and \(m\) represents the number of samples. The three indicators approaching zero means that the predicted results are closer to the measured results.

### 5. Result and Discussion

In this study, we primarily explored the performance of the MSWO model and its capability to predict future changes in ocean subsurface structures using remote sensing satellite images. Under limited training data conditions, the MSWO achieved optimal prediction accuracy. Furthermore, we compared the prediction model with the ocean subsurface structure reconstruction model. Over the 12 months of the test dataset, the prediction and reconstruction models exhibited complementary trends in model accuracy, which may offer new insights for future research.
5.1. Results of Ablation Experiment

The selection of different modules or methods can significantly impact the performance of the MSWO model. In the ablation study section, we investigated the performance of the SWO and MSWO models when incorporating global encoding and the shift window mechanism. Compared to the SWO model, the MSWO model employs the same downsampling and upsampling mechanisms as the SwinLSTM model, capturing larger-scale spatial processes by resizing the entire image. Additionally, unlike the spatial relative position encoding used in the window attention mechanism, the global encoding is introduced before the window attention layer. This global encoding allows elements within each small window to not only focus on elements within the same window but also to consider the global context. The shift window mechanism supplements the attention relationships between adjacent independent windows, aligning with the approach proposed by the Swin Transformer. The results of all ablation experiments are shown in Table 2. Overall, using down sampling to capture large-scale spatial information and introducing global spatiotemporal encoding enhances model accuracy. The best SWO model achieved an average MSE of 0.661, RMSE of 0.757, and MAE of 0.528 on the test dataset, while the best SWO-D model achieved an average MSE of 0.648, RMSE of 0.749, and MAE of 0.516 on the test dataset. For convenience, the SWO and MSWO models referred to in the following sections are the best-performing models.

Table 2. The results of ablation experiments are recorded, and the Model includes SWO and MSWO; Global represents whether global encoding of space-time was introduced, and Shifted represents whether a shifted window mechanism was used. The bold part represents the group with the best performance in the SWO and MSWO models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Global</th>
<th>Shifted</th>
<th>MSE</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWO</td>
<td>x</td>
<td>x</td>
<td>0.842 ± 0.106</td>
<td>0.835 ± 0.049</td>
<td>0.567 ± 0.026</td>
</tr>
<tr>
<td>SWO</td>
<td>√</td>
<td>x</td>
<td>0.661 ± 0.045</td>
<td>0.757 ± 0.027</td>
<td>0.528 ± 0.023</td>
</tr>
<tr>
<td>SWO</td>
<td>x</td>
<td>√</td>
<td>0.737 ± 0.051</td>
<td>0.798 ± 0.033</td>
<td>0.547 ± 0.031</td>
</tr>
<tr>
<td>SWO</td>
<td>√</td>
<td>√</td>
<td>0.788 ± 0.104</td>
<td>0.820 ± 0.051</td>
<td>0.557 ± 0.027</td>
</tr>
<tr>
<td>MSWO</td>
<td>x</td>
<td>x</td>
<td>0.737 ± 0.074</td>
<td>0.803 ± 0.047</td>
<td>0.558 ± 0.035</td>
</tr>
<tr>
<td>MSWO</td>
<td>√</td>
<td>x</td>
<td>0.648 ± 0.047</td>
<td>0.749 ± 0.026</td>
<td>0.516 ± 0.012</td>
</tr>
<tr>
<td>MSWO</td>
<td>x</td>
<td>√</td>
<td>0.804 ± 0.104</td>
<td>0.838 ± 0.058</td>
<td>0.595 ± 0.032</td>
</tr>
<tr>
<td>MSWO</td>
<td>√</td>
<td>√</td>
<td>0.676 ± 0.075</td>
<td>0.766 ± 0.044</td>
<td>0.523 ± 0.023</td>
</tr>
</tbody>
</table>

5.2. Model Comparison

In the field of spatiotemporal prediction, numerous advanced artificial intelligence models have been proposed for tasks such as video prediction and weather forecasting. Similar to these works, we utilized these established spatiotemporal sequence models as baseline models to evaluate the performance of the MSWO model. The baseline models include ConvLSTM, PredRNN, SwinLSTM, EarthFormer, SA-ConvLSTM, and SimVP. The input and output formats for MSWO were kept consistent with these baseline models, and the accuracy of the predictions was evaluated using MSE, RMSE, and MAE metrics. To further reduce the likelihood of random events, all experiments were repeated three times, with the mean and the standard deviation recorded for each experiment. Each training and testing session was conducted on an Nvidia Tesla V100 GPU. The average results obtained from all baseline models are presented in Table 3. The overall mean MSE for ConvLSTM, PredRNN, SwinLSTM, EarthFormer, SA-ConvLSTM, and SimVP, and our MSWO model was 0.829, 1.007, 1.074, 0.789, 0.935, 0.768, 0.661, and 0.648, respectively. The overall mean RMSE was 0.856, 0.925, 0.963, 0.834, 0.891, 0.813, 0.757, and 0.749, respectively. The overall mean MAE was 0.580, 0.634, 0.648, 0.596, 0.610, 0.575, 0.528, and 0.516, respectively. The MSWO model achieved the best predictive accuracy among all the compared models, indicating its advantage in predicting changes in ocean subsurface structures. This demonstrates that the MSWO model has superior predictive performance in this domain.
Table 3. Comprehensive prediction comparison results between the SWO model and other baseline models on the test dataset, where Size represents the size of the image in the calculation process, and EarthFormer and MSWO perform multi-resolution sampling operations, respectively. The bold part represents the group with the best performance in different model experiments.

<table>
<thead>
<tr>
<th>Method</th>
<th>Size (H = W)</th>
<th>MSE</th>
<th>RMSE</th>
<th>MAE</th>
<th>Cite</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvLSTM</td>
<td>16</td>
<td>0.829</td>
<td>0.858</td>
<td>0.580</td>
<td>[33]</td>
</tr>
<tr>
<td>PredRNN</td>
<td>16</td>
<td>1.007</td>
<td>0.925</td>
<td>0.634</td>
<td>[34]</td>
</tr>
<tr>
<td>SwinLSTM</td>
<td>16</td>
<td>1.074</td>
<td>0.963</td>
<td>0.648</td>
<td>[42]</td>
</tr>
<tr>
<td>EarthFormer</td>
<td>16/8</td>
<td>0.789</td>
<td>0.834</td>
<td>0.596</td>
<td>[46]</td>
</tr>
<tr>
<td>SA-ConvLSTM</td>
<td>16</td>
<td>0.935</td>
<td>0.891</td>
<td>0.610</td>
<td>[41]</td>
</tr>
<tr>
<td>SimVP</td>
<td>16</td>
<td>0.768</td>
<td>0.813</td>
<td>0.575</td>
<td>[30]</td>
</tr>
<tr>
<td>SWO (ours)</td>
<td>16</td>
<td>0.661</td>
<td>0.757</td>
<td>0.528</td>
<td>-</td>
</tr>
<tr>
<td>MSWO (ours)</td>
<td>16/8</td>
<td>0.648</td>
<td>0.749</td>
<td>0.516</td>
<td>-</td>
</tr>
</tbody>
</table>

To illustrate the variations in predictive accuracy over space and time, we have plotted Figures 5 and 6. Figure 5 presents the performance metrics of different models at various depths within the upper 250 m of the ocean, while Figure 6 shows the performance metrics across different months. All models exhibited a trend where the prediction error initially increases and then decreases, with the maximum error occurring around the 100 m depth. This corresponds to the thermocline, a transitional layer between warmer surface water and cooler deep water, characterized by a steep vertical temperature gradient. The thermocline significantly affects the ocean’s density and acoustic fields. Numerous previous studies [22,29,47,48] have highlighted the challenges artificial intelligence models face in accurately reconstructing subsurface structures due to the presence of the thermocline, posing a considerable challenge for all predictive tasks. For the study area, another challenge is the irregular variations in ocean structure caused by the El Niño–Southern Oscillation (ENSO) phenomenon. The limited availability of measured data hampers the ability of AI models to learn long-term decadal variations, thus obstructing accurate future predictions. Figure 6 shows how model errors change over time, with almost all models displaying an increase in error as the prediction horizon extends. This error increase is due to both the declining correlation between the output results and the real satellite data inputs, as well as the accumulation of errors from the recursive process. Overall, the MSWO model achieved the best predictive performance among all compared models, with the lowest error increase over time. This demonstrates that the MSWO model has broad application potential in the field of spatiotemporal sequence prediction.

Figure 5. Evaluation index results of different models at different depths. (a) MSE, (b) RMSE, (c) MAE.
5.3. Compare with P2P Schemes

In this study, we introduce a method for predicting future subsurface ocean temperature structures using historical satellite remote sensing images. However, we pose a new inquiry: How does the performance of using satellite remote sensing imagery as input for prediction compare to using historical profile data? Employing satellite remote sensing image for predicting future ocean structures is more convenient due to its lower cost of acquisition, wider coverage, and longer available period. Hence, we replaced historical satellite remote sensing data with historical measured Argo data as input, employing the MSWO model for training and testing. The experiments were replicated three times on an Nvidia Tesla V100 GPU, utilizing MSE, RMSE, and MAE as evaluation metrics. The performance of the model across depths is depicted in Figure 7.

In Figure 7, S2P (surface to profile) and P2P (profile to profile) represent predictions of future profile structures using historical satellite remote sensing and historical profile data, respectively. Unexpectedly, compared to using solely satellite remote sensing data, employing larger volumes of historical profile data did not result in a significant increase in predictive accuracy; the predictive errors of P2P remained largely comparable to those of S2P. A comparison between the S2P and P2P modes reveals the superior performance of S2P in the upper layers of the ocean and the relatively better results around 100 m depth for P2P. This suggests that variations captured by satellite remote sensing data exhibit more relevant features in the prediction process for upper ocean layers. As the inversion depth increases, the introduction of historical profile data may lead to improved inversion outcomes. Overall, the comprehensive performance of S2P and P2P modes in this experiment did not demonstrate significant differences. This finding may be attributed to the fact that changes in the upper 250 m of ocean structure are largely influenced by surface ocean dynamics, and the variations captured by historical satellite remote sensing data are already sufficiently

Figure 5. Evaluation index results of different models at different depths. (a) MSE, (b) RMSE, (c) MAE.

Figure 6. Evaluation index results of different models in the predicted 6-month period. (a) MSE, (b) RMSE, (c) MAE.
comprehensive. Despite being influenced by dataset limitations and variations related to ENSO phenomena in the study area, predictive models still have considerable room for improvement in accuracy. Nevertheless, this discovery underscores the efficacy of utilizing historical satellite remote sensing data for predicting future subsurface ocean structures, with an accuracy comparable to using historically measured data for forecasting future ocean profiles. This facilitates the direct utilization of satellite remote sensing imagery for large-scale, long-term prediction forecasts in practical applications.

![Figure 7](image-url)

**Figure 7.** Evaluation index results of historical satellite remote sensing predicting future profile structure (S2P) and historical profile predicting future profile structure (P2P) models at different depths. (a) MSE, (b) RMSE, (c) MAE.

The task of predicting the next 6 months using the preceding 24 months differs from the reconstruction task, where the input comprises the preceding 12 months’ data alongside the known 6 months’ data to reconstruct the subsurface ocean temperature structure. In contrast to future prediction tasks, reconstruction tasks benefit from the utilization of real-time satellite remote sensing data, enabling them to better capture real-time changes while learning historical patterns. To examine the similarities and differences in the inversion results between the two approaches, we conducted a comparative analysis between the prediction and reconstruction results. Specifically, we divided the test set into 12 months, which were then categorized into four seasons based on the Northern Hemisphere’s months (with December to February categorized as spring, March to May as summer, June to August as autumn, and September to November as winter), as depicted in Figure 8. Figure 8a represents the prediction results for the 12 months of 2019 obtained by different models using the prediction method, with the line graph illustrating the RMSE between predicted and actual values. Figure 8c depicts the reconstruction results for the 12 months of 2019 obtained by different models using the reconstruction method, with the line graph showing the RMSE between reconstructed and actual values. Figure 8b,d showcase the seasonal variation in errors across different models.

From Figure 8, it is evident that the prediction models generally exhibit higher errors compared to the reconstruction models. Interestingly, we observe an inverse trend in inversion errors between the prediction and reconstruction models on the test dataset (highlighted in the yellow box in the figure). We attribute this phenomenon to the prediction models primarily learning regularities on the temporal scale, lacking input of real-time spatial change data. In contrast, the reconstruction task for subsurface ocean structure better addresses this gap, optimizing inversion results further by incorporating real-time
satellite remote sensing data onto the foundation of spatiotemporal sequence prediction. This holds practical significance, particularly for real-time forecasting endeavors.

![Figures](http://example.com/figures)

**Figure 8.** (a) RMSE changes of different baseline models on the 12-month prediction task of the test dataset. (b) Seasonal RMSE bar chart of different benchmark models on the prediction task of the test dataset. (c) RMSE changes for different baseline models on 12-month reconstruction tasks of the test dataset. (d) Seasonal RMSE histogram for different benchmark models on the reconstruction task of the test dataset.

**5.4. Section Cutting Comparison**

To further validate the consistency between the three-dimensional subsurface ocean structure predicted by MSWO and the measured data, we selected four latitudinal and three longitudinal transects for cross-sectional plotting. As shown in Figure 9, along the latitude direction, we selected four transects at 15.5° N (A1), 0.5° N (A2), 15.5° S (A3), and 40.5° S (A4), while along the longitude direction, we chose three transects at 169.5° W (B1), 139.5° W (B2), and 119.5° W (B3). We present the measured Argo profiles, predicted profiles, and their differences plotted along these transects using the MSWO model on the test dataset for January 2019, as illustrated in Figures 10 and 11.

![Figures](http://example.com/figures)

**Figure 9.** Section selection diagram.

Observing the relationship between the measured and true profiles in Figure 10, it is evident that near the equator, warm water driven by the easterly winds converges into the western Pacific warm pool before sinking, leading to a gradient distribution in subsurface...
temperatures and shallower thermoclines in the eastern Pacific and deeper thermoclines in the western Pacific. Compared to the measured data, the predicted results from the MSWO model also reflect these gradient changes. Similarly, examining the relationship between the measured profiles in Figure 11, it is observed that the subsurface temperatures in the warm pool region near the equator are higher than those in the eastern Pacific, and subsurface temperatures in the low latitudes of the South Pacific are higher than those in the North Pacific. This distribution correlates strongly with the South Equatorial Current (SEC), Equatorial Under Current (EUC), North Equatorial Counter Current (NECC), and North Equatorial Current (NEC) [49], indicating that MSWO can effectively predict changes in subsurface water masses within the ocean.

![Figure 10](image10.png)
Figure 10. Profile diagram of the January 2019 test dataset drawn at A1–A4 cross sections.

![Figure 11](image11.png)
Figure 11. Profile diagram of the January 2019 test dataset drawn at B1–B3 cross sections.

5.5. Planar Error and Density Scatter

The planar error facilitates an understanding of the distribution of errors between the predicted results and the measured data, while the density scatter plot evaluates the correlation between the measured values and the model’s predicted results, as well as the robustness of the model’s performance. Thus, we generated planar error maps and density scatter plots for all baseline models in July 2019, focusing on the upper 100 m of the ocean. As depicted in Figure 12, significant errors were observed near the equator across all models, with the eastern Pacific displaying underestimations and the western Pacific...
showing overestimations, a pattern closely associated with the 2019 El Niño event. Overall, MSWO exhibited favorable performance advantages in predicting the development and changes in subsurface oceanic structures. Furthermore, compared to other baseline models, MSWO demonstrated the best robustness and most concentrated inversion effects in the density scatter plot.

Figure 12. All baseline models were compared with the MSWO model for planar error plots and density scatter plots.

6. Conclusions

With the advancement of human science, the development of artificial intelligence technology has gradually provided new methods and perspectives for realizing the strategy of transparent oceans. In this paper, we proposed the MSWO model and employed it to predict the subsurface temperature structure for the next six months in the Central South Pacific region using satellite remote sensing data from the past 24 months. MSWO introduces a global positional encoding to enrich the spatiotemporal relationship features among the data and utilizes channel attention mechanisms to extract crucial information. Subsequently, MSWO employs multiscale residual operations and window attention mechanisms to extract spatiotemporal correlations within the input data, facilitating predictions of future subsurface oceanic structures. In comparative experiments, MSWO achieved the best predictive performance, benefiting from the attention mechanism’s extraction of global spatiotemporal information and its efficient utilization. Additionally, we explored the impact of using either satellite remote sensing data alone or profile data alone as inputs on future predictions. Surprisingly, the results from both strategies were similar, indicating that we can directly predict and forecast future oceanic changes using satellite remote sensing methods, which are more practically applicable compared to profile prediction models that require scarce measured data. Furthermore, we compared the performance of MSWO in prediction tasks and reconstruction tasks. The experimental results show that the error trends of prediction models and reconstruction models exhibited complementary characteristics across the 12 months of the test dataset. The incorporation of remote sensing images as inputs in reconstruction models further complements the real-time features missing in prediction models, thereby improving inversion accuracy.
However, there are still many opportunities for further development in this research. For instance, the lack of measured datasets limits the model's ability to capture long-term decadal changes, and the monthly and spatial resolutions lead to the loss of many small and medium-scale change processes. The impact of the ENSO process on the model cannot be avoided, and the extent of its influence is currently unclear. The design of the MSWO model supports high-resolution satellite remote sensing images and large-scale spatiotemporal process forecasting. In future work, we will explore higher-resolution ocean reconstruction and forecasting.

**Author Contributions:** Conceptualization, L.W. and X.Z.; Data curation, Y.L.; Formal analysis, J.W.; Investigation, J.J.; Methodology, J.J.; Resources, X.Z.; Supervision, X.Z.; Validation, J.J., C.H. and Q.J.; Visualization, L.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Technology Support Talent Program of the Chinese Academy of Sciences (E4KY31), the Chinese Academy of Sciences pilot project (XDB42000000), the Major Science and Technology Infrastructure Maintenance and Reconstruction Project of the Chinese Academy of Sciences (DSS-WXGZ-2022), the National Key Research and Development Program (2021YFC1101504), the National Natural Science Foundation of China (42176030) and High Level Innovative Talent Project of NUDT.

**Data Availability Statement:** The Argo data can be downloaded from [http://apdrc.soest.hawaii.edu/projects/Argo/](http://apdrc.soest.hawaii.edu/projects/Argo/) (accessed on 1 May 2024); the ADT data can be downloaded from [https://www.aviso.altimetry.fr/en/data/products/](https://www.aviso.altimetry.fr/en/data/products/) (accessed on 1 May 2024); the SST data can be downloaded from [https://climatedataguide.ucar.edu/climate-data/sst-data-noaa-high-resolution-025x025-blended-analysis-daily-sst-and-ice-oisstv2](https://climatedataguide.ucar.edu/climate-data/sst-data-noaa-high-resolution-025x025-blended-analysis-daily-sst-and-ice-oisstv2) (accessed on 1 May 2024); and the SSS data can be downloaded from [https://www.catsd.fr/Products/Catalogue-CPDC/Catsd-products-from-Sextant#metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b](https://www.catsd.fr/Products/Catalogue-CPDC/Catsd-products-from-Sextant#metadata/0f02fc28-cb86-4c44-89f3-ee7df6177e7b) (accessed on 1 May 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

**References**

5. Li, J.; Sun, L.; Yang, Y.; Yan, H.; Liu, S. Upper Ocean Responses to Binary Typhoons in the Nearshore and Offshore Areas of Northern South China Sea: A Comparison Study. *Coas 2020*, *99*, 115–125. [CrossRef]
Remote Sens. 2024, 17, 2243


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.