*remote sensing*

MDPI

*Article*

# Convformer: A Model for Reconstructing Ocean Subsurface Temperature and Salinity Fields Based on Multi-Source Remote Sensing Observations

**Tao Song [1,2], Guangxu Xu [1,2], Kunlin Yang [1,2], Xin Li [1,†] and Shiqiu Peng [3,*,†]**

[1] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; tsong@upc.edu.cn (T.S.); z22070040@s.upc.edu.cn (G.X.); s21070063@s.upc.edu.cn (K.Y.); lix@upc.edu.cn (X.L.)

[2] Key Laboratory of Marine Hazards Forecasting, Ministry of Natural Resources, Beijing 100081, China

[3] State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China

[*] Correspondence: speng@scsio.ac.cn

[†] These authors contributed equally to this work.

**Abstract:** Observational data on ocean subsurface temperature and salinity are patently insufficient because in situ observations are complex and costly, while satellite remote-sensed measurements are abundant but mainly focus on sea surface data. To make up for the ocean interior data shortage and entirely use the abundant satellite data, we developed a data-driven deep learning model named Convformer to reconstruct ocean subsurface temperature and salinity fields from satellite-observed sea surface data. Convformer is designed by deeply optimizing Vision Transformer and ConvLSTM, consisting of alternating residual connections between multiple temporal and spatial attention blocks. The input variables consist of sea surface temperature (SST), sea surface salinity (SSS), sea surface height (SSH), and sea surface wind (SSW). Our results demonstrate that Convformer exhibits superior performance in estimating the temperature-salinity structure of the tropical Pacific Ocean. The all-depth average root mean square error (RMSE) of the reconstructed subsurface temperature (ST)/subsurface salinity (SS) is 0.353 °C/0.0695 PSU, with correlation coefficients ($R^2$) of 0.98663/0.99971. In the critical thermocline, although the root mean square errors of ST and SS reach 0.85 °C and 0.121 PSU, respectively, they remain smaller compared to other models. Furthermore, we assessed Convformer's performance from various perspectives. Notably, we also delved into the potential of Convformer to extract physical and dynamic information from a model mechanism perspective. Our study offers a practical approach to reconstructing the subsurface temperature and salinity fields from satellite-observed sea surface data.

**Keywords:** deep learning; ocean remote sensing; subsurface temperature (ST); subsurface salinity (SS); Transformer; physics

check for updates

**Citation:** Song, T.; Xu, G.; Yang, K.; Li, X.; Peng, S. Convformer: A Model for Reconstructing Ocean Subsurface Temperature and Salinity Fields Based on Multi-Source Remote Sensing Observations. *Remote Sens.* **2024**, *16*, 2422. https://doi.org/10.3390/rs16132422

Academic Editor: Jorge Vazquez

Received: 20 May 2024
Revised: 16 June 2024
Accepted: 27 June 2024
Published: 1 July 2024

## 1. Introduction

Accurate monitoring of ocean conditions is crucial for a comprehensive understanding of Earth's system dynamics, climate change, and marine ecosystems [1,2]. Ocean temperature and salinity are crucial parameters regulating heat transfer between the ocean and the atmosphere [3,4]. They are also closely associated with critical ocean−atmosphere thermal processes, including oceanic heatwaves [5,6], thermocline formation [7], El Niño evolution [8,9], and deep−water formation [10,11]. Moreover, research on ocean subsurface temperature and salinity fields is of great significance for understanding oceanic dynamic processes [12–14]. Therefore, accurately estimating the ocean subsurface temperature (ST) and the subsurface salinity (SS) fields is essential for understanding marine ecosystems, ocean dynamics, and climate change.

Observational data regarding subsurface temperature and salinity remains extremely limited because of the difficulty and high cost of the in situ observations [15]. Since 2004, the implementation of the Argo program has significantly enhanced global ocean observations [16]. However, its observation density, distribution, and spatial and temporal resolution need to be improved urgently [17]. In summary, we still need help accessing high-quality ocean subsurface temperature and salinity information directly.

Various remote sensing platforms and sensors have been used for ocean monitoring over the past 40 years [18], continuously providing products of many ocean parameters. Relative to the observational data, satellite-derived ocean data are copious, continuous, and have extensive spatial coverage. Although satellite observations have been limited to the ocean's surface, the parameters in the ocean's interior are dynamically related to those at the surface [19,20]. Many subsurface processes can be seen on the surface, such as internal waves, mixed layer depth, and eddies [21]. These connections between the surface and subsurface enable us to extract subsurface and deep ocean data from surface information [22–24].

Previous studies have explored two primary methods for reconstructing subsurface information using sea surface data: statistical and dynamic [25]. The typical dynamical methods are highly complex and only compelling in specific regions and conditions [26–28]. Statistical methods are more practical and flexible than dynamical approaches, making them applicable across a broader range of scenarios in the era of extensive marine data [29–31]. Simple statistical methods have limited accuracy in construction due to their inability to incorporate dynamical equations and spatiotemporal characteristics of ocean data [32–35]. In contrast, machine learning models can automatically and directly extract features and relations from data, achieving remarkable performance [36,37]. Recently, various machine learning techniques have been employed to estimate ocean interior structures, including support vector machines, random forests [38], self-organizing maps [39], artificial neural networks [40], and XGBoost 1.5.2 [41].

Furthermore, deep learning networks are characterized by deeper hidden layers and larger architectures, enabling them to capture more intricate features and have greater capacities. Deep learning techniques have recently been utilized to estimate subsurface variables over large oceanic areas. For example, Song et al. [42] utilized the convolutional long short-term memory network (ConvLSTM) to construct the subsurface temperature and salinity fields. Xie et al. [43] combined the U-net deep learning model and attention mechanism to reconstruct a high-resolution subsurface temperature field in the South China Sea. Mao et al. [44] formulated a model based on Dual Path Convolutional Neural Networks (DP-CNNs) to reconstruct ST and SS. Chen et al. [45] used long short-term memory network (LSTM) and Gaussian process regression (GPR) methods to estimate the temperature and salinity profiles in the northwest Pacific Ocean.

These approaches effectively capture nonlinear relations between surface and subsurface variables, indicating deep learning's satisfactory performance in reconstructing subsurface temperature and salinity fields from remote sensing observations. However, several issues remain:

1. Most studies' basic models are limited to CNN, ConvLSTM, and U-net. These models share the common feature of using CNN as the primary spatial feature extraction method. However, the limitations of convolutional layers, such as local receptive and fixed receptive fields, lead to each neuron's limited ability to consider information from a confined input area. This incapacity to capture longer distance dependencies between elements constrains the network's performance in processing global contextual information.
2. Most studies employ LSTM as the primary method for learning temporal information. However, LSTM still has not entirely resolved the vanishing gradient problem inherent in RNN, making it difficult to effectively capture long-range sequence dependencies, especially when dealing with global temporal information. Additionally, LSTM architectures pose challenges for parallelization.

3. Few studies have delved deeply into the physical processes underlying ocean dynamics. Some research endeavors to tackle this issue by incorporating multi-factor inputs, yet a thorough elaboration of the modeling mechanism often needs to be improved.

In recent years, the field of Natural Language Processing (NLP) has experienced a revolutionary transformation with the emergence of a self-attention-based method, Transformer [46]. In contrast to traditional CNN and LSTM, the self-attention mechanism effectively captures local and global long-range dependencies by directly comparing feature activations across all temporal and spatial positions. This exceeds the receptive field range of conventional convolutional filters, enabling a more comprehensive capture of correlated information within input sequences. Moreover, the self-attention mechanism effectively addresses the gradient vanishing and explosion issue in LSTM while exhibiting robust parallelization capabilities. Subsequently, Dosovitskiy et al. [47] proposed the Vision Transformer (ViT), which segments images into small patches, converts these patches into sequences, and then utilizes self-attention mechanisms to capture global information within the image. Building upon the abovementioned research, we propose a reconstruction architecture for subsurface temperature and salinity fields primarily consisting of alternating residual connections between multiple temporal and spatial attention blocks. The model is named Convformer, based on a deep optimization of ConvLSTM and Vision Transformer integration. The specific optimizations and modifications are detailed as follows:

1. Compared to CNN and LSTM, we utilized the Transformer's attention mechanism to extract spatiotemporal information by residually connecting spatial and temporal attention blocks. This facilitates a more comprehensive and accurate capture of spatiotemporal information on a global scale, compensating for the omission of global information from previous work.
2. The Transformer architecture inherently lacks a concept of position or sequence, necessitating the introduction of positional encoding to address this challenge. Traditional positional encoding inadequately addresses this challenge. Consequently, we employ ConvLSTM as the positional encoding layer for Convformer. ConvLSTM considers positional information in a sequential input manner, thereby employing it as a positional encoding layer facilitates effective learning of sequential spatiotemporal information. Additionally, given that the Transformer requires a large amount of data, CNN exhibits strong learning capabilities even with relatively small training sets due to its robust inductive bias. Hence, in scenarios where data is limited, employing ConvLSTM as the positional encoding layer in the model can also effectively extract spatiotemporal features at an initial stage.
3. The Vision Transformer primarily focuses on extracting spatial attention by computing attention among patches, which overlooks the internal correlations within each patch. To address this limitation, we introduce a local spatial attention mechanism that computes attention among the elements within each patch, thereby enabling a comprehensive extraction of spatial features.
4. By discussing the potential connection between residual connections and differential equations, we elucidate the Convformer's ability to capture the physical processes of ocean dynamics from a modeling mechanism.

## 2. Study Area and Data

### *2.1. Study Area*

The oceans cover 97% of the total water volume on Earth and constitute 71% of the Earth's surface area. The Pacific Ocean, the world's largest and deepest ocean, boasts the largest number of marginal seas and islands and spans across Asia, Oceania, Antarctica, and North and South America. The vast expanse of the Pacific Ocean comprises a complex and expansive aquatic system. To validate the proposed reconstruction method, we selected the central Pacific region (30°S–30°N, 160°E–120°W) as the study area, as depicted in Figure 1. This region features vast oceanic expanses and significant temperature differences, encompassing multiple oceanic current zones such as the equatorial warm current and equatorial countercurrent. It exhibits typical ocean-atmosphere interaction phenomena, including the El Niño-Southern Oscillation (ENSO). By studying this area, we can effectively verify the feasibility and effectiveness of the proposed method for reconstructing oceanic elements.

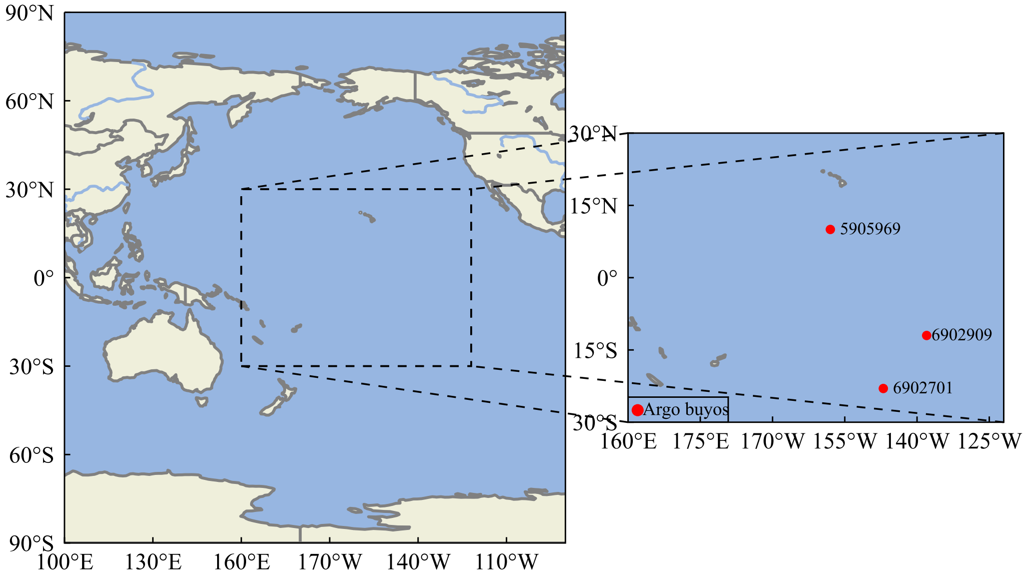

**Figure 1.** Schematic diagram and partitioning of the study area. The study area is located in the central Pacific region (160°E–120°W, 30°S–30°N); the Argo buoys No. 5905969, 6902701, and 6902909 are used for vertical profiles verification.

### 2.2. Data

In this study, we utilized two sources of ocean observational data: sea surface data obtained from satellite observations (SSS, SST, SSH, and SSW), along with gridded Argo data. Sea Surface Temperature (SST) data were obtained from the National Oceanic and Atmospheric Administration (NOAA) [48]. The dataset has provided information at a spatial resolution of 1° × 1° and a temporal resolution of 1 day. Sea Surface Salinity (SSS) data were sourced from the European Space Agency's Soil Moisture and Ocean Salinity project (SMOS), with a spatial resolution of 0.25° × 0.25° and a temporal resolution of 1 month [49]. Sea Surface Height (SSH) data were obtained from the Archiving, Validation, and Interpretation of Satellite Oceanographic datasets (AVISO), which utilizes altimeters with a spatial resolution of 0.25° × 0.25° and a temporal resolution of 7 days [50]. Sea surface wind (SSW) data were derived from the Cross-Calibration Multi-Platform Project (CCMP) with a spatial resolution of 0.25° × 0.25° and a temporal resolution of 1 month [51]. Subsurface Temperature (ST) and Subsurface Salinity (SS) were derived from the BOA_Argo data obtained from the China Argo Real-Time Data Center (CARDC), which provides a monthly globally gridded dataset of temperature and salinity profiles at 58 standard depths with a spatial resolution of 1° × 1° [52]. Specifically, we utilized training data from January 2004 through December 2017, randomly selecting 90% of the data for training and reserving the remaining 10% for validation. Finally, we evaluated the performance of the models using data from 2018, assessing performance based on the root mean square error (RMSE) and the correlation coefficient ($R^2$). To ensure that the input data are at the same spatiotemporal scale for learning and discovering data patterns, we processed all the above data into monthly averages and downsampled the spatial resolution of different datasets to a uniform 1° × 1° resolution. This operation helps to ensure that the input data have consistent and accurate spatiotemporal resolution. Given that the training data belong to different scales, it is necessary to normalize all data to the range [0, 1] for consistency before feeding them into the neural network.

## 3. Methods

### *3.1. ConvLSTM*

ConvLSTM (Convolutional Long Short-Term Memory) [53] is a type of recurrent neural network (RNN) designed specifically for spatiotemporal prediction tasks. Compared to traditional Long Short-Term Memory (LSTM) models, ConvLSTM introduces convolutional structures in both the input-to-state and state-to-state transitions, which allows the input information to be treated as a two-dimensional matrix, thus enabling a more flexible extraction of spatial information from the matrix. The critical equations of ConvLSTM are as follows. In these equations, $X_t$ represents the input at the current time step, $H_t$ is hidden state, $H_{t-1}$ is the hidden state from the previous time step, $i_t$, $f_t$, $C_t$ and $o_t$ are the activation values of the input gate, forget gate, cell state update gate, and output gate, respectively. $W$ and $b$ denote the weights and biases, with subscripts indicating their relationships with inputs and hidden states. $\sigma$ represents the sigmoid activation function, $\odot$ denotes the Hadamard product, tanh represents the hyperbolic tangent activation function, and $*$ denotes the convolution operator.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \odot C_{t-1} + b_f) \tag{2}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot C_{t-1} + bo) \tag{4}$$

$$H_t = o_t \odot \tanh(C_t) \tag{5}$$

### *3.2. Self-Attention*

The self-attention mechanism allows models to flexibly attend to information from different positions when processing sequential data. Its computational process can be represented by the following formula, where $Q$ is the query matrix, $K$ is the key matrix, $V$ is the value matrix, and $d_k$ is the dimensionality of the keys. The self-attention mechanism calculates the dot product of the query matrix with all key matrices, divides by $\sqrt{d_k}$, and applies the softmax function to obtain the weights of the values. Finally, the attention weights obtained from the softmax operation are used to compute a weighted sum of the value vectors, generating an output matrix.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

### *3.3. Multi-Head Attention*

The Multi-Head Attention mechanism extends the attention mechanism by capturing information from different relations using multiple independent attention heads. This approach enables the model to encompass a broader range of relationships. Each attention head learns specific linear transformations of queries, keys, and values. Subsequently, the outputs of these heads are concatenated and undergo a linear transformation. This process helps the model to capture different relationships and features in the sequence more comprehensively. The computation process of Multi-Head Attention is as follows: Firstly, independent linear transformations are applied to the given query matrix $Q$, key matrix $K$, and value matrix $V$, resulting in multi-head query matrices $Q_i$, multi-head key matrices $K_i$ and multi-head value matrices $V_i$, as shown in Equations (7), (8), and (9) respectively. Then, independently compute attention weights and outputs for each attention head as shown in Equation (10). Finally, concatenate the outputs of all attention heads into a large matrix and undergo a linear transformation as shown in Equation (11), where head$_i$ represents the output of the $i$-th attention head, Concat denotes the concatenation operation, and $W_o$ is the linear transformation matrix for the output.

$$Q_i = Q \cdot W_{Qi} \tag{7}$$

$$K_i = K \cdot W_{Ki} \tag{8}$$

$$V_i = V \cdot W_{Vi} \tag{9}$$

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}(\frac{Q_i {K_i}^T}{\sqrt{d_k}})V_i \tag{10}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W_o \tag{11}$$

## *3.4. Convformer*

### 3.4.1. Model Architeure

Our model overview and flowchart are depicted in Figure 2. The model consists of positional encoding, patch embedding, and an encoder, connected recursively block by block, with layer-to-layer connections within each block. The inputs are processed layer by layer to produce the final forecast. The input is fed into the ConvLSTM block to encode the positional information, enabling the model to acquire sequential information naturally and initially extract spatiotemporal features. Next, the the processed output from the ConvLSTM is split into fixed-size patches. Each of these patches undergoes linear embedding, where they are flattened and transformed into a sequence of vectors via a linear projection layer. This transformation is crucial as it prepares the spatiotemporal features for further processing by converting them into a suitable format for the encoder architecture. Finally, the resulting sequence of vectors, embedded with both spatial and temporal information, is inputted into the Convformer encoder. This encoder comprises multiple blocks of temporal and spatial attention, which iteratively refine these embeddings using self-attention mechanisms and feed-forward networks, allowing the model to capture intricate patterns and dependencies. Detailed information on each module is described as follows.

### 3.4.2. Input

The Convformer takes as input a clip $X \in \mathbb{R}^{T \times 5 \times H \times W}$ consisting of inputs for $T$ time steps, where each input comprises five channels (SST, SSS, SSH, USSW, and VSSW), with each channel having dimensions $H$ and $W$.

### 3.4.3. Positional Encodeding

For the Transformer, since all temporal sequences are processed simultaneously within the network, sequential information is lost when entering the network. Therefore, Transformers require additional processing to inform the relative position of each input. The solution is to use Positional Encoding. The traditional Positional Encoding in Transformers mainly involves mapping positional information to a fixed mathematical representation and adding it to the input embedding vector. However, this method may not be flexible enough in some cases because it encodes all positions in the same way without considering the specific semantic content of the input sequence. Thus, a practical approach is needed to thoroughly address the challenge of disregarding sequential information. LSTM considers positional information in a sequential input manner, where the input sequence is fed one element at a time. Therefore, using LSTM as a positional encoding layer can effectively capture sequential spatiotemporal information. Additionally, given that the Transformer requires a large amount of data, CNN exhibits strong learning capabilities even with relatively small training sets due to its robust inductive bias. Consequently, we employ ConvLSTM as the positional encoding layer for Convformer, which considers positional information in a sequential input manner like LSTM and effectively extracts spatiotemporal features at an initial stage due to CNN's inductive bias. The input $X \in \mathbb{R}^{T \times 5 \times H \times W}$ is first fed into the positional encoding block to extract sequential positional information and initial spatiotemporal features.

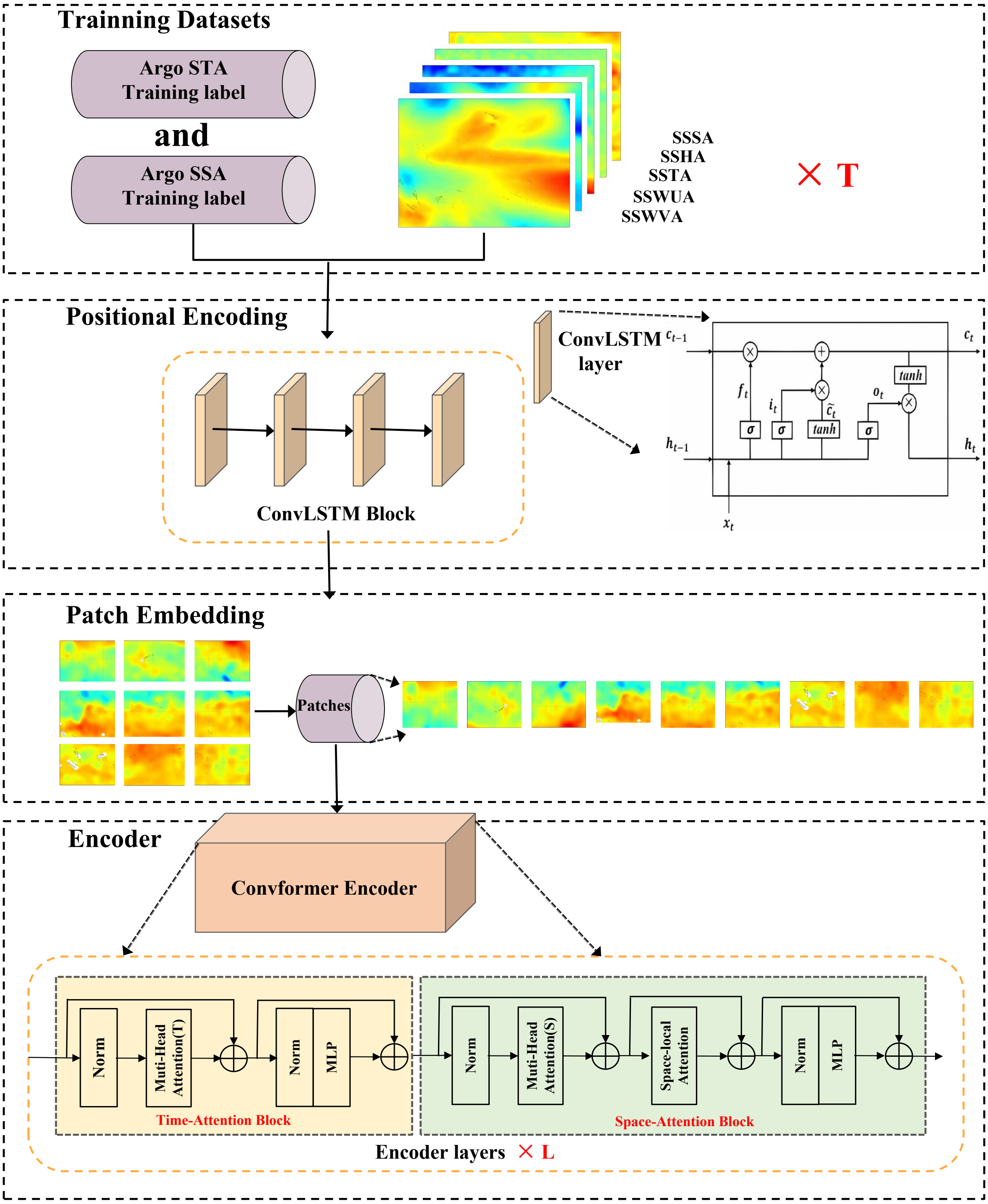

**Figure 2.** The architecture of Convformer and flowchart of Pacific ST/SS reconstruction using remote sensing data.

### 3.4.4. Patch Embeding

Following the approach of ViT [47], we decompose each channel of the processed output from the ConvLSTM into $N$ non-overlapping patches, where $N = HW/P^2$ and each patch has a size of $P \times P$. We flatten these patches into vectors $\mathbf{x}_{(n,t)} \in \mathbb{R}^{5P^2}$, where $n = 1, 2, \ldots, N$ represents spatial positions and $t = 1, 2, \ldots, T$ represents the time sequence. Then, to convert them into a suitable format for the encoder architecture, we map each patch $\mathbf{x}_{(n,t)}$ to an embedding vector $\mathbf{z}_{(n,t)} \in \mathbb{R}^D$ through a learnable linear embedding layer, as shown in Equation (12), where $W_E \in \mathbb{R}^{5P^2 \times D}$. The resulting embedding vectors $\mathbf{z}_{(n,t)}$ are fed into the encoder as input.

$$\mathbf{z}_{(n,t)} = \mathbf{x}_{(n,t)} \cdot W_E \tag{12}$$

### 3.4.5. ConvFomer Encoder

Our ConvFomer Encoder consists of $L$ encoding blocks. Each encoding block $l$ comprises a temporal attention block and a spatial attention block, both equipped with residual connections, for comprehensive extraction of temporal and spatial features, respectively. The patch embedding results are processed layer by layer. The composition of the temporal attention block and the spatial attention block is essentially similar: both comprise alternating multi-head attention mechanisms and MLP blocks, with Layer Normalization (LN) [54] applied before each block and residual connectivity applied after each block. The query/key/value vectors for multi-head attention are computed based on the representation $\mathbf{z}^{(l-1)}_{(n,t)}$ obtained from the preceding spatial or temporal attention block. The calculation process is as follows, where LN() represents LayerNorm, $h = 1, 2, \ldots, H$ is the index of multiple attention heads, and $H$ represents the total number of attention heads, with each head attention dimension set to $D_h = D/H$.

$$\mathbf{q}^{(l,h)}_{(n,t)} = W^{(l,h)}_{Q} * \mathrm{LN}(\mathbf{z}^{(l-1)}_{(n,t)}) \tag{13}$$

$$\mathbf{k}^{(l,h)}_{(n,t)} = W^{(l,h)}_{K} * \mathrm{LN}(\mathbf{z}^{(l-1)}_{(n,t)}) \tag{14}$$

$$\mathbf{v}^{(l,h)}_{(n,t)} = W^{(l,h)}_{V} * \mathrm{LN}(\mathbf{z}^{(l-1)}_{(n,t)}) \tag{15}$$

Next, the temporal attention weights $\boldsymbol{\alpha}^{(l,h)time}_{(n,t)}$ and spatial attention weights $\boldsymbol{\alpha}^{(l,h)space}_{(n,t)}$ for the query patch$_{(n,t)}$ are given by the following equations:

$$\boldsymbol{\alpha}^{(l,h)time}_{(n,t)} = \mathrm{Softmax}\left( \frac{{\mathbf{q}^{(l,h)}_{(n,t)}}^{T}}{\sqrt{Dh}} * \left[ \mathbf{k}^{(l,h)}_{(0,0)} \left\{ \mathbf{k}^{(l,h)}_{(n,t')} \right\}_{t'=1,\ldots,T} \right] \right) \tag{16}$$

$$\boldsymbol{\alpha}^{(l,h)space}_{(n,t)} = \mathrm{Softmax}\left( \frac{{\mathbf{q}^{(l,h)}_{(n,t)}}^{T}}{\sqrt{Dh}} * \left[ \mathbf{k}^{(l,h)}_{(0,0)} \left\{ \mathbf{k}^{(l,h)}_{(n',t)} \right\}_{n'=1,\ldots,N} \right] \right) \tag{17}$$

The encoding $\mathbf{z}^{l(time)}_{(n,t)}$ or $\mathbf{z}^{l(space)}_{(n,t)}$ at block $l$ is obtained by computing the weighted sum of value vectors, utilizing the self-attention coefficients of each attention head:

$$\mathbf{s}^{(l,h)(space/time)}_{(n,t)} = \sum_{t=1}^{T} \sum_{n=1}^{N} \boldsymbol{\alpha}^{(l,h)(space/time)}_{(n,t)} \mathbf{v}^{(l,h)}_{(n.t)} \tag{18}$$

Then, the vectors from all heads are concatenated together, and residual connections are applied after each operation:

$$\mathbf{z'}^{l(time/space)}_{(n,t)} = W_O \begin{bmatrix} \mathbf{s}^{(l,1)(time/space)}_{(n,t)} \\ \vdots \\ \mathbf{s}^{(l,H)(time/space)}_{(n,t)} \end{bmatrix} + \mathbf{z}^{l-1(space/time)}_{(n,t)} \tag{19}$$

It is worth mentioning that this only extracts the spatial interrelationships between the $N$ patches, while the spatial relationships within each patch are neglected. Therefore, in addition to the temporal attention block, we have further added a local-space attention mechanism to the spatial attention block to additionally extract the interactions between elements within each patch, implemented as follows:

$$\text{Attention\_}local(Q_{local}, K_{local}, V_{local}) = \text{softmax}(\frac{Q_{local}{K_{local}}^{T}}{\sqrt{d_k}})V_{local} \tag{20}$$

Finally, the results are passed through an MLP and connected using residual connections:

$$\mathbf{z}_{(n,t)}^{l(time/space)} = \text{MLP}\left(\text{LN}\left(\mathbf{z}'^{l(time/space)}_{(n,t)}\right)\right) + \mathbf{z}'^{l(time/space)}_{(n,t)} \tag{21}$$

The resulting output $\mathbf{z}_{(n,t)}^{l(time/space)}$ serves as the input for the next temporal or spatial attention block.

### 3.4.6. Residual Connections and Differential Equations

Our Convformer encoder applies residual connectivity and layer normalization both between temporal and spatial attention blocks as well as between their internal self-attention blocks and feedforward network blocks. This is because residual networks have a subtle relationship with differential equations. Each residual connection block can be represented as:

$$y_{t+1} = y_t + \text{G}(\text{LN}(y_t)) \tag{22}$$

where $y_t$ represents the output at position $t$, LN() represents the layer normalization function, and G() represents the computation function of the current layer, such as the self-attention layer or the feed-forward layer. Iterative updating can be interpreted as a discretization of a continuous function transformation. For simplicity, $\text{G}(\text{LN}(y_t))$ can be represented as the function. Then, if we relax $y_t$ to be a continuous function $y(t)$, we can rewrite Equation (22) as (23):

$$y(t + \Delta t) = y(t) + \Delta t F(y(t)) \tag{23}$$

where $\Delta t$ represents the change of $t$, also known as the step size. We can use the limit to adjust $\Delta t$ to obtain the following equation:

$$\lim_{\Delta t \to 0} \frac{y(t + \Delta t) - y(t)}{\Delta t} = F(y(t)) \tag{24}$$

Thus, we can infer that each residual connection block can be viewed as describing a first-order differential equation. Moreover, based on the formulation of the Covformer block in (22)–(24) and the universal approximation theorem of neural networks, the complexity of the network can be increased to model higher-order differential equations. Therefore, we employ residual connections between each block, enabling the model to learn to simulate underlying dynamic equations on its own, thereby enhancing the physical interpretability of the model.

## 4. Results

### *4.1. Convformer Performance against Typical Models*

To evaluate the performance of our model, we conducted subsurface temperature and salinity field reconstruction experiments using the same dataset with the ConvLSTM, ViT (Vision Transformer), and the popular Attention U-Net models in related research. We calculated the Root Mean Square Error (RMSE) and the correlation coefficient (R²) between the temperature and salinity profiles obtained from these models and the Argo profiles to assess the performance of the different methods.

Figure 3 displays the vertical distribution of RMSE for temperature profiles estimated by different models. The models exhibit relatively small RMSE on the sea surface. However, the RMSEs increased sharply with depth, reaching their maximum at 150 m depth, then sharply decreased with two subsequent increases at 500 m and 1100 m depth, respectively. This phenomenon may be attributed to the complex dynamical processes in the ocean's upper layers and perturbations in the mixing and thermocline layers, making the temperature distribution at specific depths exceptionally complex while deep-sea water remains relatively stable. Compared to other methods, Convformer demonstrates significantly superior performance in ST estimation, achieving the smallest RMSE at all depth levels. Particularly noteworthy is that at depths where temperature changes sharply (such as 100 m, 500 m, 1100 m, etc.), Convformer's RMSE is significantly smaller than that of other models. This suggests that Convformer possesses more substantial potential to approximate highly nonlinear functions effectively. This is because at these depths, where temperature changes rapidly, the relationship between temperature and depth becomes more complex, potentially involving more intricate dynamic processes and significant nonlinear characteristics.

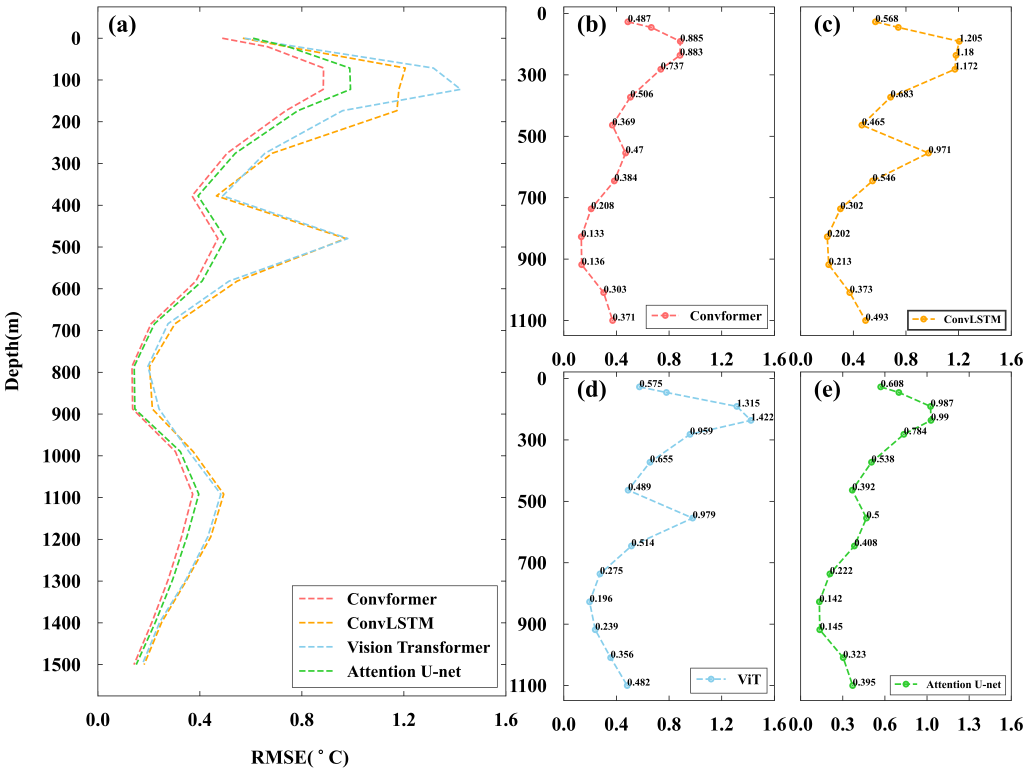

**Figure 3.** Vertical RMSE profiles for each model from 50 m down to 1500 m of ST: (**a**) models comparison; (**b**–**e**) each model respectively.

Figure 4 depicts the bar graph of ST reconstruction using different models. The overall RMSEs and correlation coefficients at different depth levels between the estimates of each model and ARGO grid data are summarized in Table 1. Overall, all models showed good performance with similar trends. The models' average RMSE and correlation coefficients were 0.486/0.489/0.386/0.353 and 98.253%/97.795%/98.213%/98.663%, respectively. The RMSE of the Convformer model is significantly smaller than that of CovLSTM and ViT at all depth layers, with mean values reduced by 27.36% and 27.98%, respectively. Moreover, Convformer's estimated ST profiles exhibit higher correlation coefficients. At depths close to the sea surface, such as 30 m and 50 m, all models exhibit relatively low RMSE values, possibly because they are closer to the input temperature observed at the surface. However, Convformer still demonstrates significantly lower RMSE, implying its more robust feature extraction capability in accurately capturing temperature variations near the sea surface. Furthermore, Convformer may excel at modeling complex nonlinear relationships near the sea surface, as various factors such as solar radiation and wind speed influence the sea surface temperature.

**Table 1.** RMSE and correlation coefficient ($R^2$) of ST between ConvLSTM, Vision Transformer, Attention U-NET and Convformer. Bolded fonts represent the optimal results among the models.

| Depth/m | RMSE/°C | | | | $R^2$/% | | | |
|---|---|---|---|---|---|---|---|---|
| | ConvLSTM | ViT | A-U-Net | Convformer | ConvLSTM | ViT | A-U-Net | Convformer |
| 30 | 0.568 | 0.575 | 0.608 | **0.487** | 98.432 | 98.339 | 98.114 | **99.071** |
| 50 | 0.742 | 0.780 | 0.746 | **0.665** | 97.546 | 97.110 | 97.352 | **98.158** |
| 100 | 1.205 | 1.315 | 0.987 | **0.885** | 96.740 | 95.083 | 97.135 | **97.525** |
| 150 | 1.180 | 1.422 | 0.990 | **0.883** | 96.862 | 94.305 | 97.377 | **97.845** |
| 200 | 1.172 | 0.959 | 0.784 | **0.737** | 97.919 | 96.974 | 98.017 | **98.357** |
| 300 | 0.683 | 0.655 | 0.538 | **0.506** | 98.578 | 98.378 | 98.551 | **98.950** |
| 400 | 0.465 | 0.489 | 0.392 | **0.369** | 97.633 | 97.206 | 97.551 | **97.926** |
| 500 | 0.971 | 0.979 | 0.500 | **0.470** | 96.560 | 95.304 | 95.559 | **96.947** |
| 600 | 0.546 | 0.514 | 0.408 | **0.384** | 97.275 | 96.724 | 97.155 | **97.650** |
| 700 | 0.302 | 0.275 | 0.222 | **0.208** | 97.850 | 97.361 | 97.912 | **98.186** |
| 800 | 0.202 | 0.196 | 0.142 | **0.133** | 98.072 | 97.698 | 97.960 | **98.370** |
| 900 | 0.213 | 0.239 | 0.145 | **0.136** | 98.309 | 98.009 | 98.272 | **98.604** |
| 1000 | 0.373 | 0.356 | 0.323 | **0.303** | 98.220 | 98.067 | 98.259 | **98.578** |
| 1100 | 0.493 | 0.482 | 0.395 | **0.371** | 98.409 | 98.269 | 98.446 | **98.720** |
| 1200 | 0.442 | 0.431 | 0.349 | **0.328** | 98.473 | 98.371 | 98.613 | **98.792** |
| 1300 | 0.349 | 0.346 | 0.293 | **0.275** | 98.687 | 98.591 | 98.886 | **98.983** |
| 1400 | 0.250 | 0.242 | 0.225 | **0.212** | 98.877 | 98.768 | 98.680 | **99.188** |
| 1500 | 0.181 | 0.173 | 0.150 | **0.141** | 99.154 | 99.087 | 99.066 | **99.470** |
| 1600 | 0.144 | 0.139 | 0.122 | **0.115** | 99.370 | 99.302 | 99.255 | **99.669** |
| 1700 | 0.105 | 0.104 | 0.090 | **0.085** | 99.457 | 99.404 | 99.402 | **99.773** |
| 1800 | 0.058 | 0.053 | 0.052 | **0.045** | 99.553 | 99.523 | 99.508 | **99.865** |
| 1900 | 0.038 | 0.037 | 0.037 | **0.033** | 99.589 | 99.620 | 99.618 | **99.952** |
| Average | 0.486 | 0.489 | 0.386 | **0.353** | 98.253 | 97.795 | 98.213 | **98.663** |

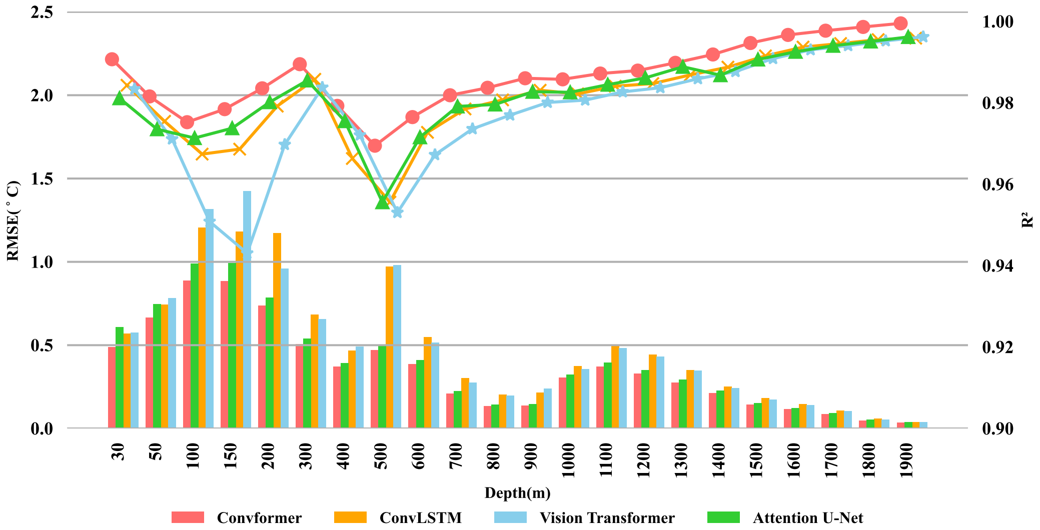

**Figure 4.** Bar graph for each model from 50 m down to 1900 m of ST. The legend colors correspond to various models, while the symbols indicate the correlation coefficients' values for each model at a given depth.

At depths of 100 m and 150 m, the RMSE of each model reaches its maximum value, possibly due to the involvement of complex dynamical processes and temperature variations at these depths. However, Convformer exhibits a much more moderate increase than the other models, and the gap in RMSE between Convformer and other models reaches its peak. This suggests that Convformer may excel at capturing complex dynamical processes and the intricate nonlinear relationships of temperature variations, demonstrating outstanding physical modeling capabilities. Subsequently, the RMSE of each model sharply decreases, with two increases occurring at depths of 500 m and 1100 m, respectively. However, the increase in RMSE for Convformer remains relatively moderate. In the depth range of 1100 m to 1900 m, the RMSE of each model shows a sharp decrease, and the differences between them are slight. This indicates that all models perform excellently in the relatively stable deep water layers.

Figure 5 shows the vertical distribution of RMSE for salinity profiles estimated by different models. The trend of increase and decrease in salinity profiles is almost identical to temperature, indicating that the distribution at specific depths in this region becomes exceptionally complex due to dynamic changes and substantial thermocline variability. It is evident from the figure that Convformer continues to demonstrate excellent performance in salinity estimation, particularly at depths where salinity changes drastically (e.g., 150 m, 500 m, 1100 m, etc.). At these depths, the RMSE of the salinity profiles estimated by Convformer is significantly smaller than that of other models, further highlighting the model's robust nonlinear approximation and physical modeling capabilities.

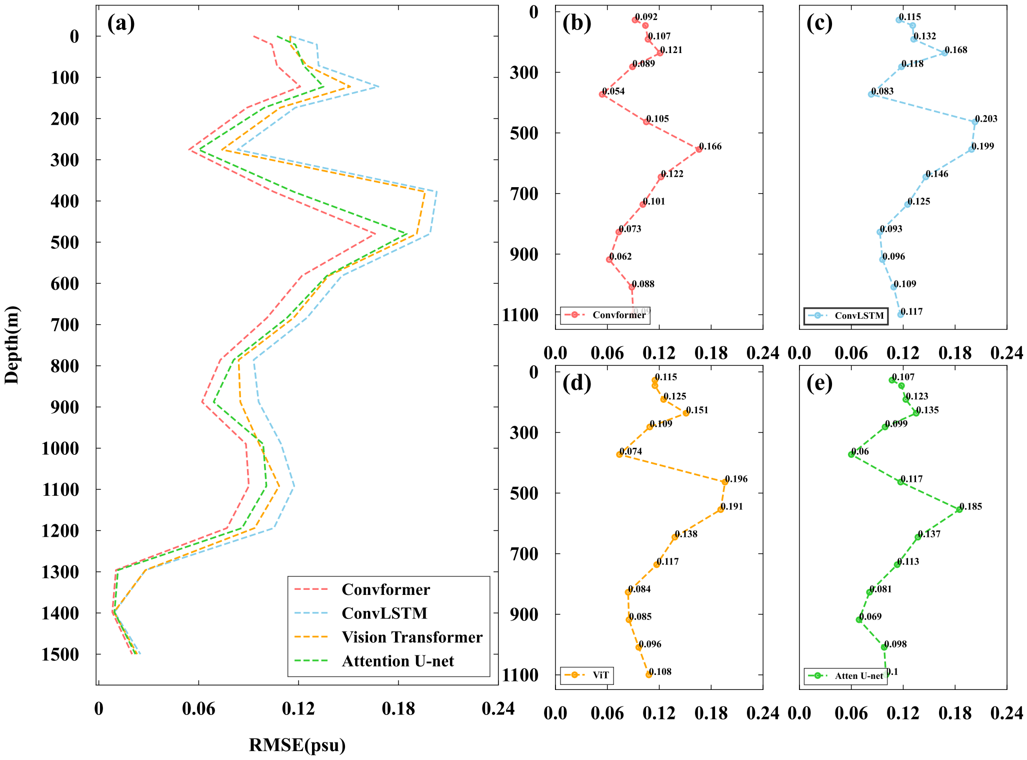

**Figure 5.** Vertical RMSE profiles for each model from 50 m down to 1500 m of SS: (**a**) models comparison; (**b**–**e**) each model respectively.

Figure 6 illustrates the bar graph of SS reconstruction using different models. The overall RMSEs and correlation coefficients at various depth levels, depicting the estimates of each model compared to ARGO grid data, are summarized in Table 2. Notably, the errors of salinity reconstruction models are significantly more minor than those of temperature reconstruction models. The average RMSE and correlation coefficients for each model are 0.09327/0.08649/0.07814/0.06951 and 99.96900%/99.96691%/99.96746%/99.97145%, respectively.

**Table 2.** RMSE and correlation coefficient ($R^2$) of SS between ConvLSTM, Vision Transformer, Attention U-NET and Covformer. Bolded fonts represent the optimal results among the models.

| Depth/m | RMSE/psu | | | | $R^2$/% | | | |
|---|---|---|---|---|---|---|---|---|
| | ConvLSTM | ViT | A-U-Net | Convformer | ConvLSTM | ViT | A-U-Net | Convformer |
| 30 | 0.115 | 0.115 | 0.107 | **0.092** | 99.90794 | 99.89066 | 99.89220 | **99.91870** |
| 50 | 0.131 | 0.115 | 0.118 | **0.104** | 99.86417 | 99.86597 | 99.87158 | **99.87522** |
| 150 | 0.168 | 0.151 | 0.135 | **0.121** | 99.87167 | 99.86614 | 99.86373 | **99.87265** |
| 200 | 0.118 | 0.109 | 0.099 | **0.089** | 99.90785 | 99.90153 | 99.91057 | **99.91341** |
| 300 | 0.083 | 0.074 | 0.060 | **0.054** | 99.96621 | 99.96467 | 99.96565 | **99.96757** |
| 400 | 0.203 | 0.196 | 0.117 | **0.105** | 99.97482 | 99.97107 | 99.94515 | **99.97500** |
| 500 | 0.199 | 0.191 | 0.185 | **0.166** | 99.98963 | 99.98773 | 99.98899 | **99.98978** |
| 600 | 0.146 | 0.138 | 0.137 | **0.122** | 99.99631 | 99.99527 | 99.99601 | **99.99665** |
| 700 | 0.125 | 0.117 | 0.113 | **0.101** | 99.99741 | 99.99623 | 99.99710 | **99.99773** |
| 800 | 0.093 | 0.084 | 0.081 | **0.073** | 99.99810 | 99.99692 | 99.99762 | **99.99831** |
| 900 | 0.096 | 0.085 | 0.069 | **0.062** | 99.99848 | 99.99748 | 99.99789 | **99.99851** |
| 1000 | 0.109 | 0.096 | 0.098 | **0.088** | 99.99870 | 99.99789 | 99.99846 | **99.99880** |
| 1100 | 0.117 | 0.108 | 0.100 | **0.090** | 99.99912 | 99.99849 | 99.99825 | **99.99915** |
| 1200 | 0.105 | 0.093 | 0.086 | **0.077** | 99.99932 | 99.99860 | 99.99690 | **99.99933** |
| 1300 | 0.028 | 0.028 | 0.011 | **0.010** | 99.99984 | 99.99984 | 99.99973 | **99.99985** |
| 1400 | 0.009 | 0.009 | 0.009 | **0.008** | 99.99986 | 99.99985 | 99.99984 | **99.99987** |
| 1500 | 0.025 | 0.023 | 0.022 | **0.020** | 99.99990 | 99.99989 | 99.99988 | **99.99990** |
| 1600 | 0.015 | 0.013 | 0.014 | **0.012** | 99.99992 | 99.99991 | 99.99990 | **99.99992** |
| 1700 | 0.010 | 0.009 | 0.009 | **0.008** | 99.99993 | 99.99992 | 99.99991 | **99.99993** |
| 1800 | 0.008 | 0.007 | 0.007 | **0.006** | 99.99994 | 99.99994 | 99.99994 | **99.99995** |
| 1900 | 0.006 | 0.006 | 0.006 | **0.005** | 99.99995 | 99.99995 | 99.99995 | **99.99996** |
| Average | 0.093 | 0.086 | 0.078 | **0.069** | 99.96900 | 99.96691 | 99.96746 | **99.97145** |

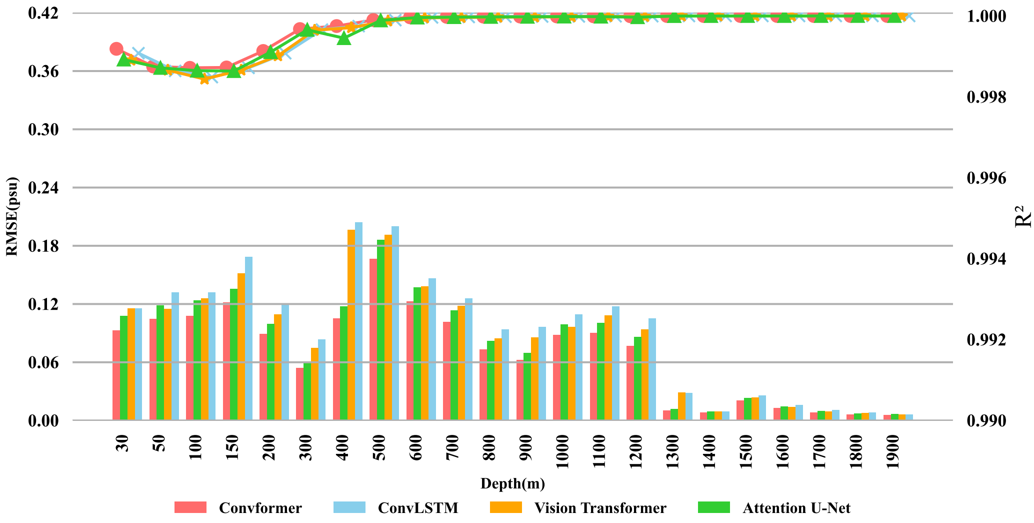

**Figure 6.** Bar graph for each model from 50 m down to 1900 m of SS. The legend colors correspond to various models, while the symbols indicate the correlation coefficients' values for each model at a given depth.

At each depth level, the salinity reconstruction RMSE of the Convformer model remains significantly smaller than that of CovLSTM, ViT, and Attention U-net, with reductions of 25.47%, 19.63%, and 11.04%, respectively. However, it is noteworthy that the reduction in RMSE for salinity is relatively minor. This is primarily due to the overall more straightforward structure of salinity, which makes all models perform superiorly and accurately on salinity compared to temperature, resulting in less room for Convformer to improve. However, within 30 m to 50 m, the RMSE of salinity for Convformer remains a significant estimation ability at this depth level. This highlights Convformer's capability to better capture surface spatiotemporal features. At 400 m and 500 m depths, the salinity RMSE of all models reaches its maximum value. However, Convformer still performs better than other models, particularly at 400 m, where it achieves reductions of 48.33% and 46.34% relative to ConvLSTM and ViT, respectively. At depths where RMSE briefly increases, such as at 1000 m and 1100 m, Convformer's accuracy improves significantly. This further highlights the model's outstanding capability in capturing complex deep-sea dynamic processes and nonlinear relationships, which is particularly suitable for extracting more complex salinity distribution features. In the depth range of 1300 m to 1900 m, the accuracy of all models is very high; however, while Convformer's RMSE remains lower than the other models, the difference is insignificant. This suggests that within this depth range, the salinity distribution is relatively simple, and the features are easy to extract, resulting in relatively minor performance improvements.

Overall, in the experiments reconstructing subsurface temperature and salinity fields, Convformer demonstrates superior performance compared to models such as ConvLSTM, ViT, and Attention U-net. Its average RMSE shows a significant percentage reduction, accompanied by higher correlation coefficients, further validating Convformer's superiority in reconstructing subsurface temperature and salinity fields. Of particular note is Convformer's strong capability in capturing complex deep-sea dynamic processes and modeling nonlinear relationships within depth ranges where such processes are intricate and nonlinear relationships predominate.

### *4.2. Sensitivity Analysis of Model Input*

To assess the sensitivity of temperature and salinity profile estimation to various input parameters, we conducted a detailed comparison of Convformer's estimates under different training input conditions.

Figure 7 displays the vertical distribution of the root mean square error (RMSE) of ST and SS under various input conditions, with detailed results recorded in Tables 3 and 4. Observing the root mean square temperature error, we noticed that differences are small but sufficient to account for differences in sensitivity between input elements.

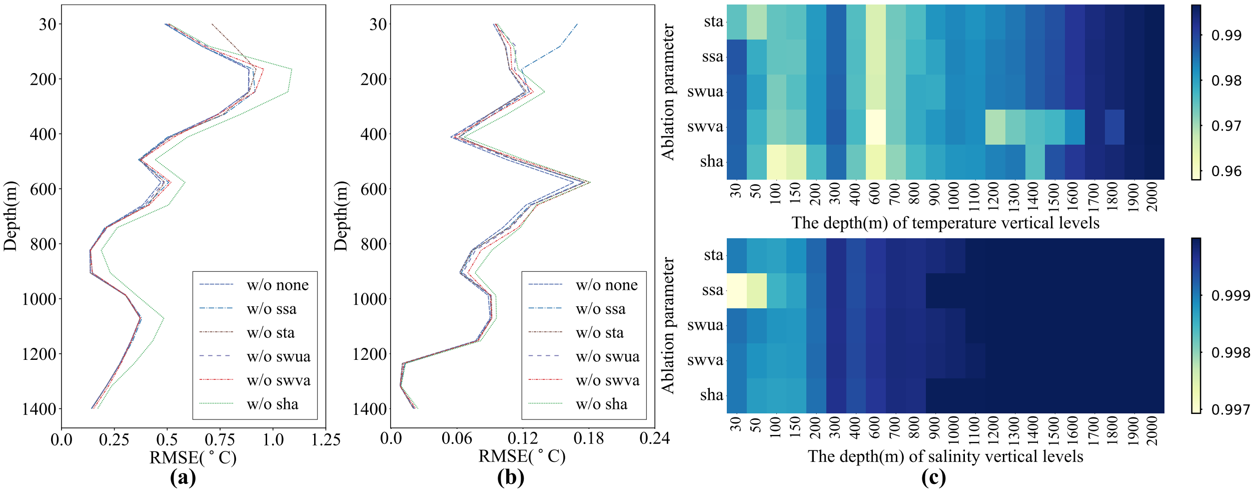

**Figure 7.** Vertical RMSE profiles for Convformer with different remote sensing inputs schemes from 50 m down to 1500 m: (**a**) RMSE of ST; (**b**) RMSE of SS; (**c**) heat map of correlation coefficients.

**Table 3.** RMSE for SS From Convformer With Different Inputs. The bold represents the optimal.

| Depth/m | RMSE/psu | | | | | |
|---|---|---|---|---|---|---|
| | w/o SSA | w/o STA | w/o SWUA | w/o SWVA | w/o SWHA | w/o None |
| 30 | 0.169 | 0.095 | 0.094 | 0.094 | 0.096 | **0.092** |
| 50 | 0.153 | 0.103 | 0.113 | 0.108 | 0.111 | **0.104** |
| 100 | 0.118 | 0.107 | 0.112 | 0.109 | 0.115 | **0.107** |
| 150 | 0.125 | 0.123 | 0.125 | 0.129 | 0.139 | **0.121** |
| 200 | 0.092 | 0.088 | 0.089 | 0.093 | 0.105 | **0.089** |
| 300 | 0.061 | 0.061 | 0.056 | 0.057 | 0.066 | **0.054** |
| 400 | 0.111 | 0.112 | 0.111 | 0.117 | 0.121 | **0.105** |
| 500 | 0.174 | 0.176 | 0.175 | 0.181 | 0.181 | **0.166** |
| 600 | 0.126 | 0.127 | 0.127 | 0.133 | 0.131 | **0.122** |
| 700 | 0.106 | 0.107 | 0.109 | 0.115 | 0.117 | **0.101** |
| 800 | 0.074 | 0.074 | 0.076 | 0.081 | 0.091 | **0.073** |
| 900 | 0.063 | 0.062 | 0.065 | 0.069 | 0.076 | **0.062** |
| 1000 | 0.090 | 0.091 | 0.090 | 0.090 | 0.095 | **0.088** |
| 1100 | 0.091 | 0.091 | 0.090 | 0.091 | 0.095 | **0.090** |
| 1200 | 0.078 | 0.078 | 0.078 | 0.079 | 0.081 | **0.077** |
| 1300 | 0.012 | 0.010 | 0.012 | 0.011 | 0.013 | **0.010** |
| 1400 | 0.008 | 0.008 | 0.008 | 0.008 | 0.009 | **0.008** |
| 1500 | 0.020 | 0.020 | 0.021 | 0.022 | 0.024 | **0.020** |
| 1600 | 0.013 | 0.012 | 0.013 | 0.013 | 0.015 | **0.012** |
| 1700 | 0.009 | 0.008 | 0.008 | 0.008 | 0.009 | **0.008** |
| 1800 | 0.006 | 0.006 | 0.006 | 0.006 | 0.007 | **0.006** |
| 1900 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 | **0.005** |
| Average | 0.078 | 0.071 | 0.072 | 0.074 | 0.077 | **0.069** |

**Table 4.** RMSE for ST from Convformer with different inputs. The bold represents the optimal.

| Depth/m | RMSE/°C | | | | | |
|---|---|---|---|---|---|---|
| | w/o SSA | w/o STA | w/o SWUA | w/o SWVA | w/o SWHA | w/o None |
| 30 | 0.492 | 0.711 | 0.496 | 0.506 | 0.510 | **0.487** |
| 50 | 0.666 | 0.819 | 0.691 | 0.678 | 0.701 | **0.665** |
| 100 | 0.905 | 0.920 | 0.887 | 0.956 | 1.089 | **0.885** |
| 150 | 0.915 | 0.888 | 0.887 | 0.915 | 1.071 | **0.883** |
| 200 | 0.770 | 0.764 | 0.738 | 0.741 | 0.846 | **0.737** |
| 300 | 0.502 | 0.517 | 0.512 | 0.532 | 0.596 | **0.506** |
| 400 | 0.364 | 0.370 | 0.376 | 0.377 | 0.443 | **0.369** |
| 500 | 0.507 | 0.488 | 0.484 | 0.518 | 0.584 | **0.470** |
| 600 | 0.402 | 0.411 | 0.394 | 0.415 | 0.506 | **0.384** |
| 700 | 0.205 | 0.214 | 0.217 | 0.212 | 0.264 | **0.208** |
| 800 | 0.137 | 0.138 | 0.135 | 0.137 | 0.188 | **0.133** |
| 900 | 0.146 | 0.136 | 0.137 | 0.146 | 0.231 | **0.136** |
| 1000 | 0.305 | 0.304 | 0.308 | 0.307 | 0.356 | **0.303** |
| 1100 | 0.380 | 0.372 | 0.375 | 0.373 | 0.484 | **0.371** |
| 1200 | 0.337 | 0.332 | 0.330 | 0.330 | 0.432 | **0.328** |
| 1300 | 0.278 | 0.278 | 0.278 | 0.282 | 0.344 | **0.275** |
| 1400 | 0.213 | 0.212 | 0.212 | 0.222 | 0.237 | **0.212** |
| 1500 | 0.141 | 0.144 | 0.141 | 0.152 | 0.171 | **0.141** |
| 1600 | 0.115 | 0.116 | 0.115 | 0.115 | 0.138 | **0.115** |
| 1700 | 0.086 | 0.087 | 0.090 | 0.089 | 0.100 | **0.085** |
| 1800 | 0.048 | 0.047 | 0.044 | 0.047 | 0.053 | **0.045** |
| 1900 | 0.033 | 0.033 | 0.033 | 0.033 | 0.035 | **0.033** |
| Average | 0.361 | 0.377 | 0.358 | 0.367 | 0.426 | **0.353** |

As for ST, significant errors across all depths when SSHA information was not included in the input conditions. This may be because the interaction of multiple factors on the subsurface, including heat exchange, internal thermal expansion, and ocean circulation, can lead to significant changes in sea surface height. The upper ocean (<150 m) demonstrates the highest RMSE without SSTA input. However, beyond a depth of 150 m, the disparity between the model without SSTA input and the one utilizing all parameters as input is insignificant. This suggests that SSTA significantly affects model performance in the upper ocean, while its impact diminishes in deeper waters. SSTA may be linked to more intricate dynamic processes and temperature fluctuations in the upper ocean, causing significant RMSE when SSTA input is absent, as the model fails to capture these complexities accurately. Conversely, in deeper waters, other parameters may begin to govern temperature dynamics, reducing the relative influence of SSTA. Consequently, the performance gap between the model without SSTA input and the one utilizing all parameters as input diminishes in deeper layers. Moreover, the RMSE from the model without SSA indicates an insignificant correlation between SSSA and ST, suggesting a weak association between SSSA and the complex dynamic variations of subsurface temperature. This could be attributed to the independent dynamic processes experienced by SSSA and temperature profiles within the ocean interior, where surface salinity anomalies do not directly influence temperature changes. The impact of SWUA and SWVA in the upper ocean (<150 m) resembles that of SSTA, showing significant RMSE. This is likely due to sea surface wind being one of the primary driving forces behind surface ocean water movement. Consequently, variations in sea surface wind may directly influence surface ocean flow, leading to a higher correlation between SWUA, SWVA, and temperature profiles in surface regions. It is worth noting that the RMSE of without SWVA is relatively large, indicating that the model is more sensitive to variations in SWVA. This may reflect the model's higher sensitivity to vertical wind shear in the ocean. Additionally, at a depth of 500 m, without SWVA and SWUA, all input modes exhibit the maximum RMSE for temperature. This is likely because, at this depth, the dynamic processes within the ocean become more complex, involving more vertical movements and interactions between water masses. SWVA and SWUA are typically associated with seawater's vertical and horizontal movement, and their absence may lead to inadequate dynamic capture of temperature changes at this depth level.

The relationship between salinity and temperature on density is a fundamental aspect influencing SSH changes, resulting in a specific correlation between SSHA and salinity profiles. Therefore, similar to temperature, salinity profiles estimated at all depths under input conditions with SSHA information exhibit significant errors. SSTA does not notably enhance the reconstruction of salinity profiles. However, within the depth range of 400 m to 600 m, the RMSE of salinity profiles in the model without SSTA input is noticeably larger. This indicates that temperature contributes to the estimation of salinity profiles to some extent in water regions with more complex dynamic processes and salinity distributions. In the model without SSSA input, the upper ocean (<150 m) demonstrates larger RMSE values. Compared to the impact of SSTA on temperature profiles, the error in salinity profiles in the upper ocean is more prominent, indicating a higher sensitivity to SSSA in this region. This sensitivity may stem from SSSA often reflecting deviations of sea surface salinity from long-term averages, which could be closely related to upper ocean dynamic processes and surface salinity distributions. The impacts of SWUA and SWVA on salinity are almost analogous to those on temperature. In the upper ocean and the relatively complex 400–600 m structure, errors are prominent, and still more significant errors in the model without SWVA. This further indicates that the surface ocean flow, as well as the vertical movement and interaction of water masses induced by sea surface wind, significantly influence the distribution of both salinity and temperature.

This study highlights the critical role of sea surface parameters in improving model accuracy, particularly in the upper layers of the ocean. Among these parameters, the SSHA emerged as a pivotal factor in estimating subsurface conditions.

### *4.3. Effects of Different Components of Convformer*

We conducted ablation studies to assess the influence of the time attention block, global space attention block, local space attention block, and residual connections on Convfomer's performance. As temperature exhibits more significant and pronounced variations compared to salinity, we chose subsurface temperature as an example for the ablation experiments to facilitate a more precise comparison of the results.

Figure 8 illustrates the vertical distribution of RMSE of temperature profiles estimated by different combinations, with detailed results documented in Table 5. The time attention block in the Convformer model plays a crucial role in handling temporal dimension information, as it can capture dependencies and evolutionary patterns between different time steps in time series data. Therefore, ablating the time attention block may result in the model not fully utilizing information on the temporal dimension. From the results, ablating the time attention block leads to an overall increase in RMSE values of temperature profiles. The average RMSE increases from 0.353 to 0.399, representing a 13% increase. The global space attention block captures spatial dependencies between each patch. Ablating this block may limit the model in handling global information, thus affecting its understanding and prediction capabilities for the overall data. Similar to ablating the time attention block, ablating the global space attention block also increases RMSE values of temperature profiles. The increase is particularly significant at greater depths, with the average RMSE being 0.406, representing a 15% increase. The local space attention block is responsible for capturing spatial dependencies within patches; ablating this module may result in the model's inability to effectively identify and utilize local information, thereby reducing its ability to grasp detailed information. The results indicate that after ablating the local space attention block, there is a slight increase in the RMSE values of temperature profiles compared to ablating the global space attention block. This implies that while the contribution of the local space attention block to model performance is relatively minor, it still constitutes a part of the overall model performance. Residual connections alleviate the vanishing gradient and exploding gradient problems, facilitating better information propagation within the network. According to previous analysis, residual connections can simulate physical differential equations. Hence, ablating residual connections may lead to difficulties in information propagation, impacting the model's ability for physical modeling and subsequently influencing its optimization capability and performance. Experimental results indicate that ablating residual connections results in a significant increase in RMSE values of temperature profiles, with errors reaching their maximum, further affirming the importance of residual connections in the Convformer model.

In summary, the time attention block, global space attention block, local space attention block, and residual connections play crucial roles in the performance of the Convformer model. Ablating any of these modules can result in a pronounced decline. This highlights the necessity and effectiveness of these modules in the Convformer model and their significance in the subsurface temperature and salinity field reconstruction experiments.

**Table 5.** RMSE for ST from models with different components in ablation Study. "w/o T" represent Convformer without time attention block, "w/o S" represent Convformer without global space attention block, "w/o S_local" represent convformer without local space attention block and "w/o residual" represent convformer without residual connection. The bold represent the best.

| **Depth/m** | **RMSE/°C** | | | | |
|---|---|---|---|---|---|
| | **w/o T** | **w/o S** | **w/o S_Local** | **w/o Residual** | **Convformer** |
| 30 | 0.506 | 0.502 | 0.501 | 0.667 | **0.487** |
| 50 | 0.648 | 0.657 | 0.664 | 0.747 | **0.665** |
| 100 | 0.964 | 1.014 | 0.967 | 1.104 | **0.885** |
| 150 | 0.978 | 0.968 | 0.992 | 1.071 | **0.883** |
| 200 | 0.794 | 0.828 | 0.851 | 0.837 | **0.737** |

**Table 5.** *Cont.*

| Depth/m | RMSE/°C | | | | |
|---|---|---|---|---|---|
| | w/o T | w/o S | w/o S_Local | w/o Residual | Convformer |
| 300 | 0.562 | 0.575 | 0.561 | 0.608 | **0.506** |
| 400 | 0.419 | 0.433 | 0.409 | 0.434 | **0.369** |
| 500 | 0.512 | 0.562 | 0.531 | 0.589 | **0.470** |
| 600 | 0.470 | 0.484 | 0.473 | 0.454 | **0.384** |
| 700 | 0.245 | 0.234 | 0.248 | 0.258 | **0.208** |
| 800 | 0.172 | 0.173 | 0.165 | 0.165 | **0.133** |
| 900 | 0.215 | 0.206 | 0.205 | 0.210 | **0.136** |
| 1000 | 0.348 | 0.349 | 0.348 | 0.358 | **0.303** |
| 1100 | 0.467 | 0.468 | 0.471 | 0.476 | **0.371** |
| 1200 | 0.422 | 0.424 | 0.424 | 0.435 | **0.328** |
| 1300 | 0.337 | 0.331 | 0.339 | 0.348 | **0.275** |
| 1400 | 0.233 | 0.230 | 0.236 | 0.240 | **0.212** |
| 1500 | 0.168 | 0.166 | 0.165 | 0.178 | **0.141** |
| 1600 | 0.129 | 0.134 | 0.132 | 0.145 | **0.115** |
| 1700 | 0.098 | 0.099 | 0.096 | 0.100 | **0.085** |
| 1800 | 0.049 | 0.051 | 0.049 | 0.052 | **0.045** |
| 1900 | 0.034 | 0.034 | 0.035 | 0.035 | **0.033** |
| Average | 0.399 | 0.406 | 0.403 | 0.432 | **0.353** |

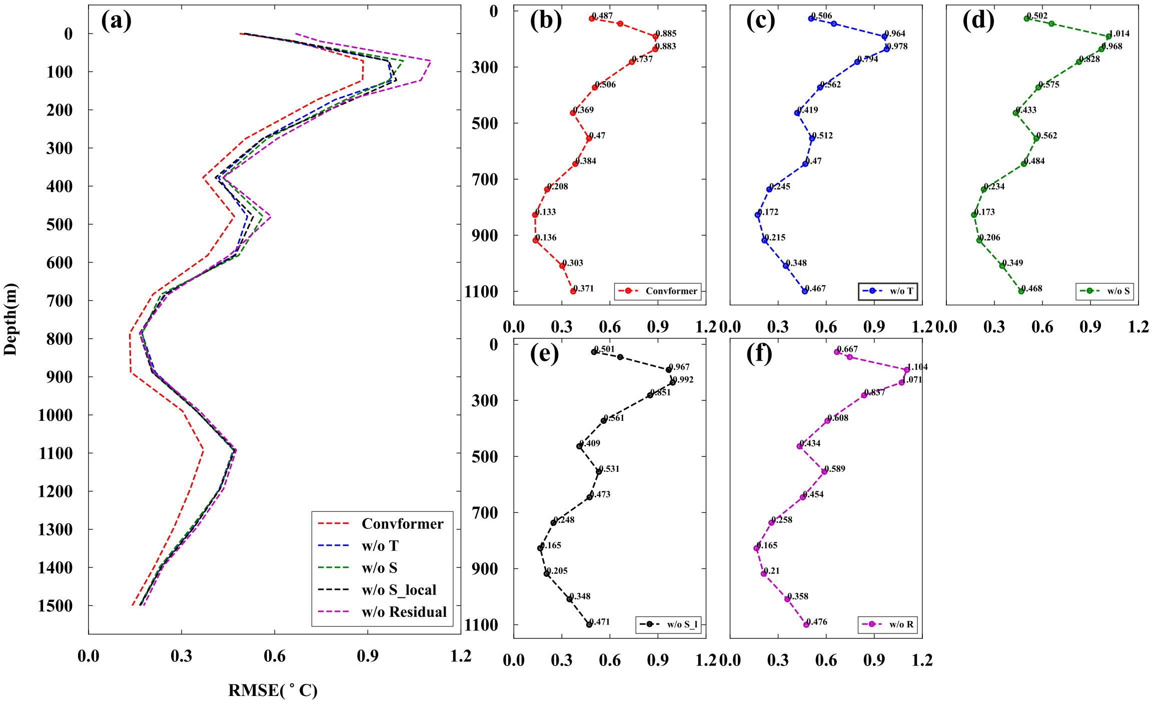

**Figure 8.** Vertical RMSE profiles for Convformer with different components in ablation study from 50 m down to 1500 m of ST: (**a**) models comparison; (**b**–**f**) each model respectively.

## *4.4. Error Analysis of Reconstruction Result*

### 4.4.1. Temporal Error Analysis

The RMSE for temperature and salinity reconstructed using the Convfomer model for different months in 2018 is depicted in the Figure 9 and Figure 10 respectively. Each grid point represents the average measure between predicted and actual values for that month at the specified depth, with all performance results obtained from the test set.

Overall, the temperature reconstruction experiment shows considerable differences in RMSE between months at shallower depths (30–600 m), whereas the variations between months at deeper levels (700–1900 m) are less pronounced. This is probably because the shallow waters are affected by more factors such as wind, insolation, and seasonal mixing, which can lead to rapid changes in temperature. In contrast, deeper waters are relatively stable and less affected by surface climate variations, resulting in more minor differences in prediction errors between months at these depth levels. The errors between the layers are almost always more significant from February to May and September to November. This is mainly due to the occurrence of El Niño phenomena during these periods, resulting in abnormally high ST. These abnormal sea temperatures can affect atmospheric circulation, thereby influencing the temperature distribution of the ocean's surface layer.

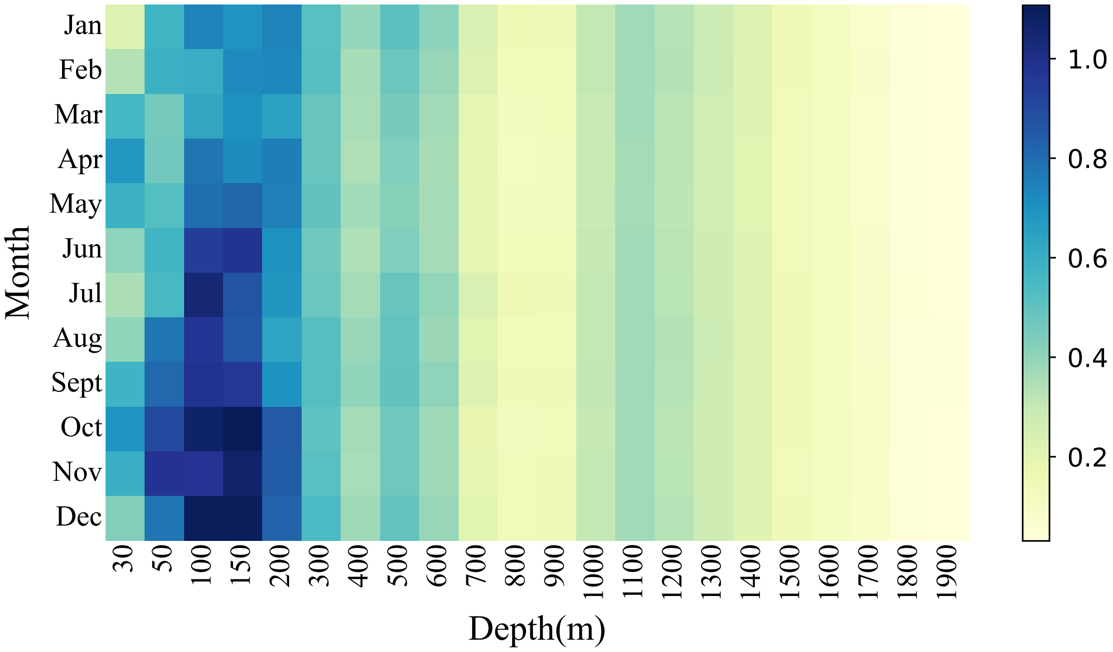

**Figure 9.** RMSE of the ST estimation at each depth level and in each month of 2018.

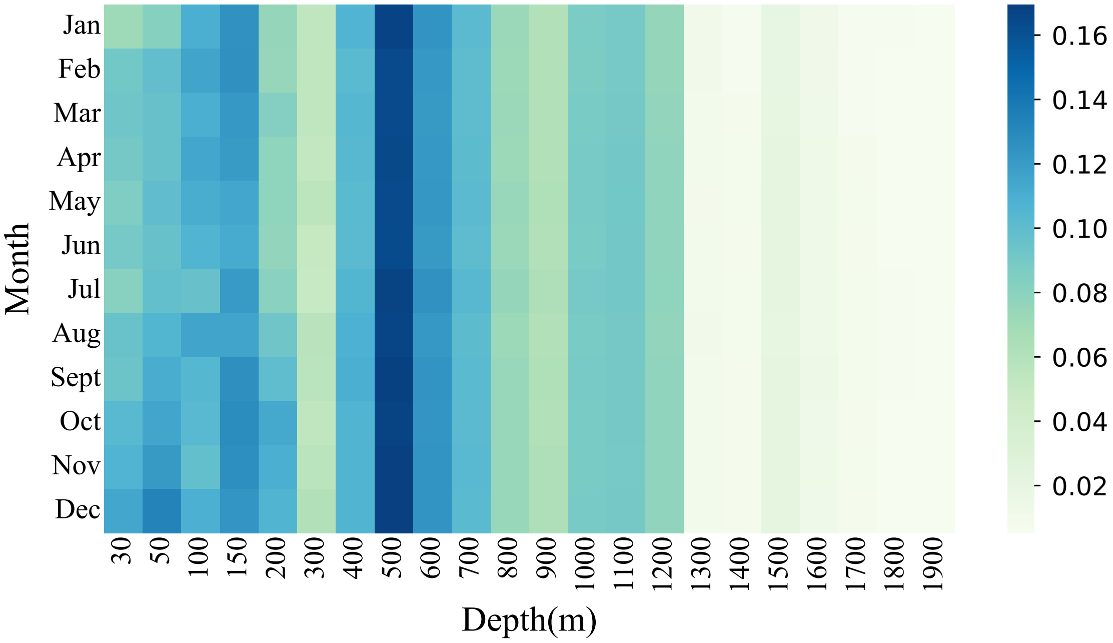

**Figure 10.** RMSE of the SS estimation at each depth level and in each month of 2018.

The variations in errors between months for salinity are relatively minor compared to temperature. This is partly because the salinity distribution itself is more stable than temperature, and it also underscores the robust learning ability of our model in capturing spatiotemporal variations in salinity. Salinity errors are generally higher in autumn and winter compared to spring and summer. This is mainly due to the seasonal solid mixing and convection that typically occur during spring and summer in the ocean, resulting in higher salinity uniformity within the water column. In autumn and winter, however, the sea undergoes stratification and becomes more stagnant, forming distinct water masses and interfaces. This increases spatial variability and complexity in salinity, making it challenging for the model to estimate salinity distribution, resulting in increased errors accurately. Overall, our model can accurately estimate subsurface temperature and salinity field distribution with consistent performance across varying depths and times.

### 4.4.2. Longitude Profile Validation

To deepen our understanding of the distribution of subsurface temperature and salinity fields in the Pacific Ocean and assess the reconstruction performance of the Conformer model from a vertical perspective, we chose a profile for analyzing vertical variations in ST and SS.

Figure 11 depicts a vertical profile along 188.5° longitude. Figure 11a shows widespread warm water areas in the central and western equatorial Pacific due to a weak La Niña phenomenon, with the southern areas exhibiting warmer temperatures compared to the northern regions. Surface seawater from the eastern Pacific is pushed towards the central western Pacific, leading to a noticeable warm water zone below 200 m in depth. The La Niña effect gradually diminishes as depth increases, and the ST distribution stabilizes beyond 300 m. In the Pacific region, around 5–10°N, there is a noticeable cold zone from approximately 50 m to around 400 m depth, likely caused by accelerated flow of the North Equatorial Warm Current during La Niña. Figure 11c shows the reconstruction results of the Conformer model for the vertical profile of ST at longitude 188.5° (171.5°W). The reconstruction results of the model closely resemble those of the Argo observation grid, with 95.86% of prediction errors within ±1.2 °C and 85.5% within ±0.8 °C. Figure 11b illustrates the vertical profile of Argo SS at longitude 188.5°. Salinity distribution is similar to temperature, with higher values in the southern part than in the northern part, gradually decreasing with depth. A high-salinity water mass is present between 100 and 250 m depth, a feature less prominent in the temperature profile. Figure 11d presents the SS profile reconstructed by the Conformer model at longitude 188.5°. Overall, 99.49% of profile points exhibit errors within ±0.3 practical salinity units (PSU), with over 91.59% showing errors within ±0.2 PSU. Thus, it can be concluded that the Conformer model's reconstruction results are excellent. The overall predictive trends of temperature and salinity profiles resemble the actual values, indicating the robust performance of our model.

### 4.4.3. Spatial Error Analysis

Figures 12 and 13 compare the subsurface temperature and salinity fields reconstructed by the Convformer model at specific periods and the corresponding Argo labels at different depths (50, 100, 300, 600, and 1000 m). Regarding spatial distribution, the subsurface temperature and salinity fields estimated by the Convformer model demonstrate a spatial distribution pattern consistent with the Argo grid data.

The ST distribution estimated by the Convformer model shows its high consistency with Argo data at various depths. The Convformer model accurately captures the sea surface data in the target area and demonstrates notable temperature characteristics. At a depth of 50 m, both the Convformer model and Argo data indicate a trend of higher temperatures in the western Pacific Ocean compared to the eastern Pacific Ocean. As one moves from the equator towards the poles, sea water temperature gradually decreases, forming significant temperature fronts near 20°N and 24°S. The temperature differences in most regions range from −0.5 °C to 0.5 °C, with primary errors occurring in the eastern equatorial Pacific region, likely attributed to the influence of the La Niña phenomenon. The Convformer model's ST estimation at a depth of 100 m exhibits good consistency with Argo observations. However, the reconstructed details gradually fade in the marginal areas, and the contours become smoother. Compared to a depth of 50 m, the differences at 100 m have slightly increased, ranging from −0.8 °C to 0.8 °C. It is worth noting that relatively significant differences are observed in the region between the equator and 10°N, which may be attributed to the presence of the thermocline and the influence of upwelling currents. At a depth of 300 m, contrary to the shallow seas, both the Convformer model and Argo data demonstrate that sea water temperature in the equatorial region is notably lower than on both sides, with differences more minor than those at 100 m depth. Furthermore, the error in the western Pacific Ocean exceeds that in the eastern Pacific Ocean, likely attributed to the influence of ocean circulation and climate change. With increasing depth, temperature becomes more stable. At depths of 500 m and below, the differences between temperature estimates from the Convformer model and temperature values from Argo data are relatively small, ranging from −0.1 °C to 0.1 °C.

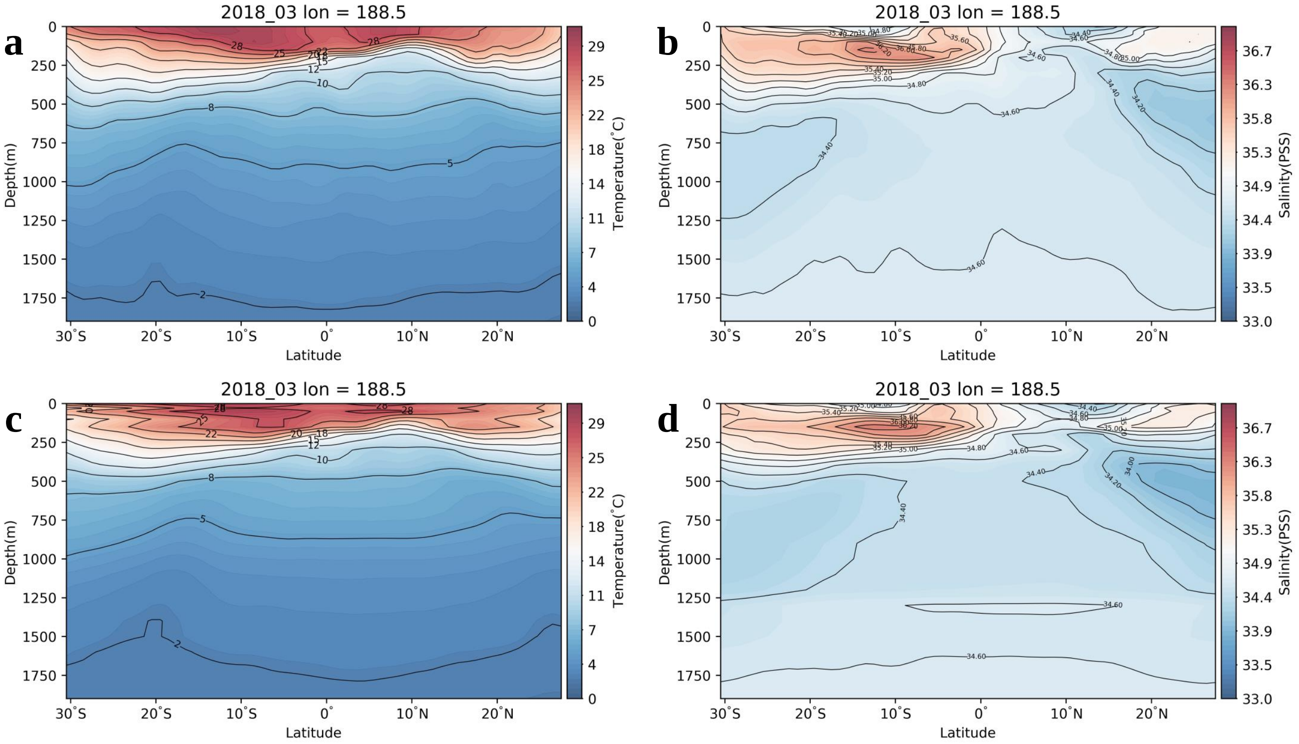

**Figure 11.** The vertical profiles of the ocean subsurface temperature and salinity fields in March 2018 at the longitude of 188.5° for (**a**) Argo ST, (**b**) Argo SS, (**c**) Convformer-reconstructed ST, and (**d**) Convformer-reconstructed SS.

These findings suggest that the Convformer model accurately estimates subsurface temperatures in the tropical Pacific. Similarly, the Convformer model exhibits good consistency with Argo gridded data in estimating SS, with no significant differences observed between the Convformer model's estimates and Argo data, effectively capturing the distribution characteristics of SS. In the upper ocean (50 m and 100 m), salinity differences in most regions range from −0.24 psu to 0.24 psu. With increasing depth, salinity becomes more stable, and the differences between salinity estimates from the Convformer model and those derived from Argo data are less than 0.1 psu. These results affirm the accuracy of the Convformer model in estimating salinity in the tropical Pacific.

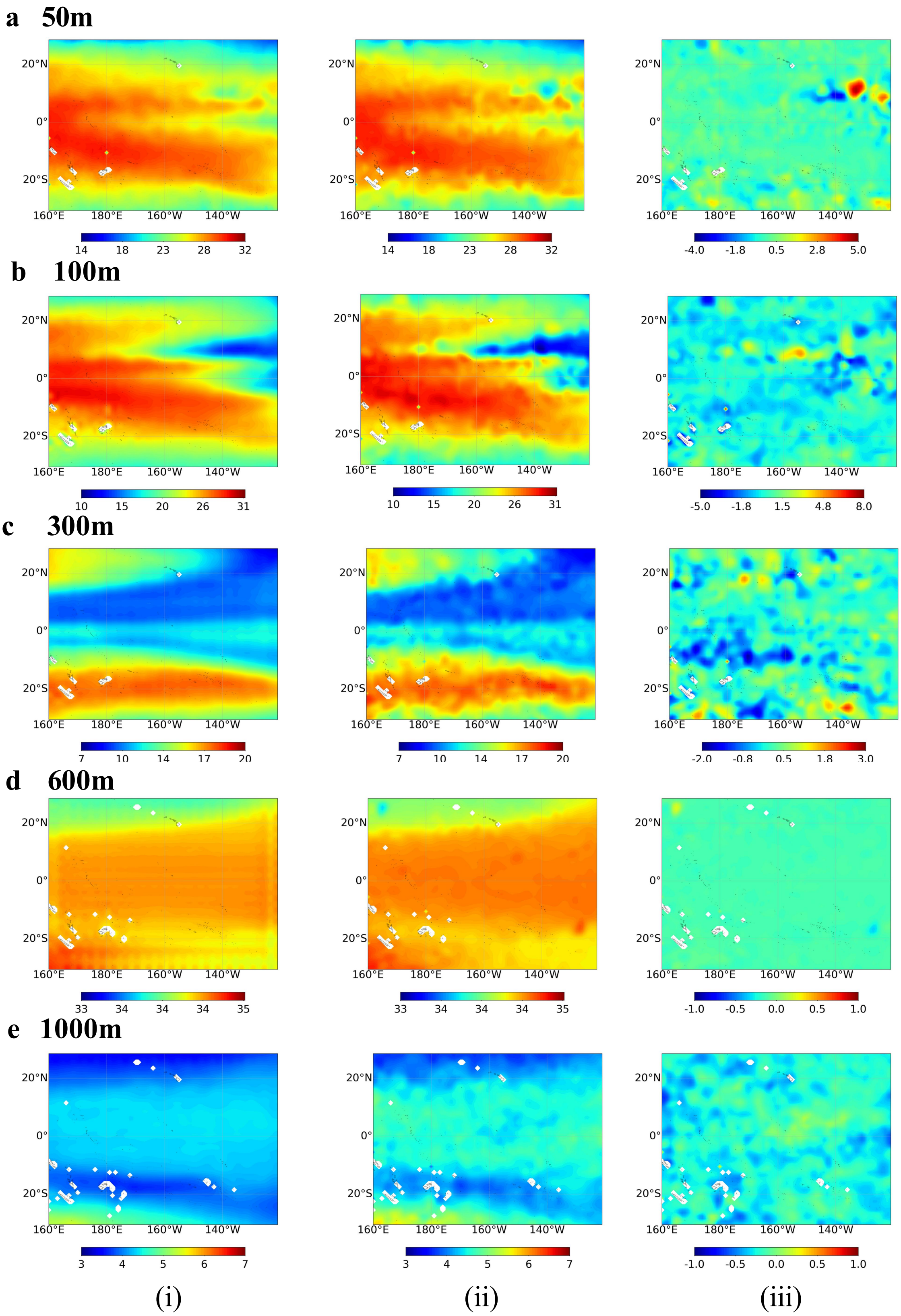

**Figure 12.** Distribution of temperature field estimated by the (**i**) Convformer compared to the (**ii**) Argo ST and their (**iii**) difference (dS = STConvfomer − STArgo) at the depth of (**a**) 50 m, (**b**) 100 m, (**c**) 300 m, (**d**) 600 m, and (**e**) 1000 m.

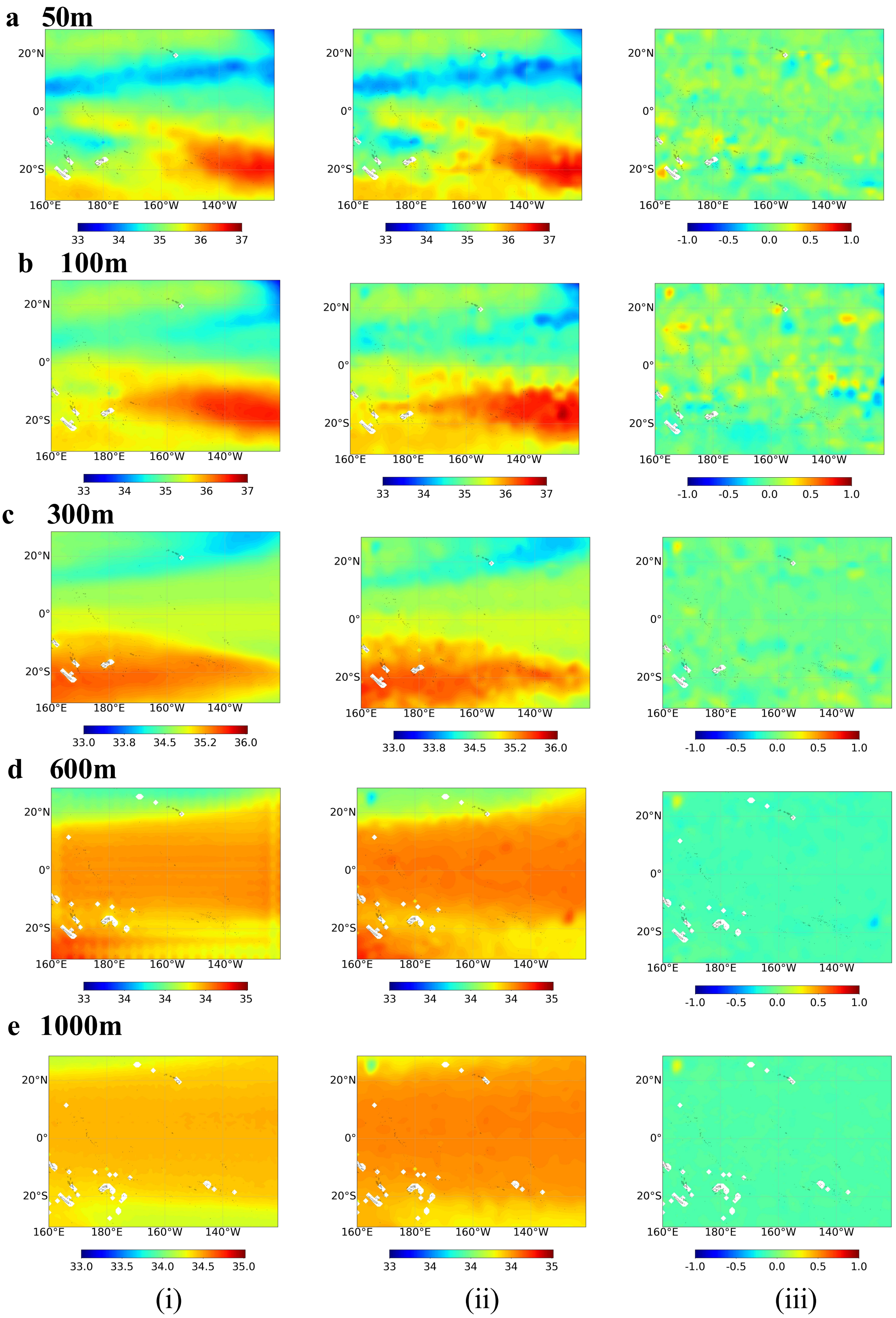

**Figure 13.** Distribution of salinity field estimated by the (**i**) Convformer compared to the (**ii**) Argo ST and their (**iii**) difference (dS = SSConvfomer − SSArgo) at the depth of (**a**) 50 m, (**b**) 100 m, (**c**) 300 m, (**d**) 600 m, and (**e**) 1000 m.

## 5. Discussion

The Convformer model represents a significant advancement in estimating subsurface temperature and salinity in the tropical Pacific Ocean, showcasing notable improvements over existing models such as ConvLSTM, Vision Transformer, and Attention U-Net. This enhanced performance is primarily attributed to the unique spatiotemporal attention mechanism and residual connections integrated within the Convformer architecture, facilitating the effective extraction and representation of complex oceanographic processes. At depths near the thermocline (100–150 m), the accuracy of temperature and salinity predictions decreases. However, the model still achieves minor errors compared to other models. This demonstrates the robustness of Convformer to extract complex spatial and temporal features.

We comprehensively assess the model in terms of model comparison, ablation experiments, temporal errors, spatial errors, and longitude profiles. We also explored the potential of Convformer to extract physical and dynamic information from a model mechanism perspective. Together, these demonstrate that our study offers a practical approach to reconstructing ST and SS from satellite-observed sea surface data.

The comprehensive assessment conducted in this study underscores the reliability of the Convformer model in reconstructing subsurface temperature and salinity fields across various depths. This is critical for applying deep learning methods in oceanographic research and provides valuable insights for enhancing future ocean models.

## 6. Conclusions

In marine science research, accurately estimating the temperature-salinity structure of the ocean subsurface is crucial for a deeper understanding of ocean dynamics and climate change. We propose a noval neural network model called Convformer to tackle this challenge. This model integrates satellite remote sensing data and observational data to reconstruct the subsurface temperature and salinity fields of the tropical Pacific. We utilized sea surface elements, including sea surface temperature, sea surface salinity, sea surface height, and sea surface wind, to reconstruct subsurface temperature and salinity fields at various depths. Argo gridded products and float profiles were employed as experimental labels and validation sets. The results indicate that the Convformer model excels in the task of reconstructing temperature-salinity fields, surpassing models like ConvLSTM, Vision Transformer, and Attention U-Net, as evidenced by smaller RMSE and more significant correlation coefficients values. This superior performance is likely attributed to the Convformer model's unique spatiotemporal attention mechanism and the potential for extracting physical information through residual connections, allowing for better capture of spatiotemporal information and representation of complex ocean processes.

Additionally, the study conducted a comprehensive assessment of the performance of the Convformer model from various angles. We evaluated the influence of sea surface parameters on the performance of the Convformer model, revealing that sea surface variables are more significant in the upper ocean, especially in shallow waters. Sea surface height anomaly (SSHA) was identified as one of the most critical factors for estimating subsurface temperature and salinity. Additionally, we compared the model-estimated subsurface temperature and salinity with Argo gridded data at different depths. The results demonstrate that the Convformer model exhibits robust performance, effectively capturing the characteristics of the observed subsurface temperature and salinity fields. However, due to the presence of a thermocline, the accuracy of temperature and salinity estimates decreases at depths around 100–150 m, presenting a challenge for accurate estimation. Nonetheless, the Convformer model performs admirably overall, accurately estimating ST and SS below 500 m.

In summary, the proposed Convformer model demonstrates exceptional performance in estimating ST and SS in the tropical Pacific. The model exhibits high accuracy and reliability in reconstructing subsurface temperature and salinity at various depths. The findings of this study are crucial for applying machine learning methods to subsurface temperature and salinity fields reconstruction and offer insights for the improvement and development of future ocean models. However, as a statistical tool, the Convformer model has limitations in estimating extreme anomaly events. In future research, we will utilize more advanced machine learning methods and further integrate ocean dynamic mechanisms. This will help improve the robustness and accuracy of the model, thereby enhancing our understanding and prediction of extreme events in the ocean. Since 2016, some global salinity datasets have experienced severe drift, and we will attempt to address the drift problem with this method. Additionally, the model's applicability can be expanded to vast oceanic regions, allowing for precise reconstruction of subsurface temperature and salinity fields in multiple oceanic areas and even the gradual establishment of global products for subsurface temperature and salinity. This can contribute to practical applications such as sound propagation, mixed layer depth (MLD) estimation, and ocean disaster prediction. Additionally, the model can be extended to estimate other critical ocean parameters, such as velocity fields and ocean density, offering extensive exploration opportunities for future research. By broadening the application scope of the Convformer model, we can gain a deeper understanding of the complexity of the ocean system and provide more avenues for exploring ocean dynamics and climate change.

**Author Contributions:** Conceptualization, T.S. and G.X.; methodology, T.S. and S.P.; software, T.S., G.X. and K.Y.; validation, G.X., S.P. and X.L.; formal analysis, T.S. and S.P.; investigation, G.X. and K.Y.; resources, T.S. and X.L.; data curation, G.X. and K.Y.; writing—original draft preparation, T.S., G.X. and S.P.; writing—review and editing, G.X. and K.Y.; visualization, G.X. and X.L.; supervision, T.S. and S.P.; project administration, T.S. and S.P.; funding acquisition, T.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Major Projects of National Natural Science Foundation of China (U20A20105), National Key Research and Development Project of China (2021YFA1000103, 2021YFA1000102), National Natural Science Foundation of China (Grant Nos. 61972416, 62272479, 62202498), Taishan Scholarship (tsqn201812029), Shandong Provincial Natural Science Foundation(ZR2021QF023), Fundamental Research Funds for the Central Universities (21CX06018A), Spanish project PID2019-106960GB-I00, Juan de la Cierva IJC2018-038539-I.

**Data Availability Statement:** Sea Surface Temperature (SST) data are obtained from the National Oceanic and Atmospheric Administration (NOAA, http://apdrc.soest.hawaii.edu/data/data.php, accessed on 10 April 2023). Sea Surface Salinity (SSS) data are sourced from the European Space Agency's Soil Moisture and Ocean Salinity project (SMOS, http://eopi.esa.int, accessed on 10 April 2023). Sea Surface Height (SSH) data are obtained from the Archiving, Validation, and Interpretation of Satellite Oceanographic datasets (AVISO, http://www.aviso.altimetry.fr, accessed on 16 April 2023). Sea surface wind (SSW) data are derived from the Cross-Calibration Multi-Platform Project (CCMP, https://rda.ucar.edu/datasets/ds745.1/, accessed on 8 May 2023). The BOA_Argo data obtained by accessing the China Argo Real-Time Data Center (CARDC, http://www.argo.org.cn, accessed on 10 April 2023).

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

## References

1. Stewart, R.H. *Introduction to Physical Oceanography*; Texas A&M University: College Station, TX, USA, 2004.
2. Bindoff, N.L.; Cheung, W.W.; Kairo, J.G.; Arístegui, J.; Guinder, V.A.; Hallberg, R.; Hilmi, N.J.M.; Jiao, N.; Karim, M.S.; Levin, L.; et al. Changing ocean, marine ecosystems, and dependent communities. In *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*; Cambridge University Press: Cambridge, UK, 2019 ; pp. 477–587.
3. Trenberth, K.E.; Fasullo, J.T. An apparent hiatus in global warming? *Earth's Future* **2013**, *1*, 19–32. [CrossRef]
4. Johnson, G.C.; Lyman, J.M. Warming trends increasingly dominate global ocean. *Nat. Clim. Chang.* **2020**, *10*, 757–761. [CrossRef]
5. Pearce, A.F.; Feng, M. The rise and fall of the "marine heat wave" off Western Australia during the summer of 2010/2011. *J. Mar. Syst.* **2013**, *111*, 139–156. [CrossRef]
6. Oliver, E.C.; Donat, M.G.; Burrows, M.T.; Moore, P.J.; Smale, D.A.; Alexander, L.V.; Benthuysen, J.A.; Feng, M.; Sen Gupta, A.; Hobday, A.J.; et al. Longer and more frequent marine heatwaves over the past century. *Nat. Commun.* **2018**, *9*, 1324. [CrossRef]
7. Chen, Z.; Zhou, T.; Zhang, L.; Chen, X.; Zhang, W.; Jiang, J. Global land monsoon precipitation changes in CMIP6 projections. *Geophys. Res. Lett.* **2020**, *47*, e2019GL086902. [CrossRef]
8. Wallace, J.; Rasmusson, E.; Mitchell, T.; Kousky, V.; Sarachik, E.; Von Storch, H. On the structure and evolution of ENSO-related climate variability in the tropical Pacific: Lessons from TOGA. *J. Geophys. Res. Ocean.* **1998**, *103*, 14241–14259. [CrossRef]

9. Planton, Y.Y.; Vialard, J.; Guilyardi, E.; Lengaigne, M.; McPhaden, M.J. The asymmetric influence of ocean heat content on ENSO predictability in the CNRM-CM5 coupled general circulation model. *J. Clim.* **2021**, *34*, 5775–5793. [CrossRef]
10. Sprintall, J.; Tomczak, M. On the formation of Central Water and thermocline ventilation in the southern hemisphere. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **1993**, *40*, 827–848. [CrossRef]
11. Qi, J.; Qu, T.; Yin, B.; Chi, J. Variability of the South Pacific western subtropical mode water and its relationship with ENSO during the Argo period. *J. Geophys. Res. Ocean.* **2020**, *125*, e2020JC016134. [CrossRef]
12. Chen, X.; Tung, K.K. Varying planetary heat sink led to global-warming slowdown and acceleration. *Science* **2014**, *345*, 897–903. [CrossRef]
13. Klemas, V.; Yan, X.H. Subsurface and deeper ocean remote sensing from satellites: An overview and new results. *Prog. Oceanogr.* **2014**, *122*, 1–9. [CrossRef]
14. Meng, L.; Zhuang, W.; Zhang, W.; Yan, C.; Yan, X.H. Variability of the shallow overturning circulation in the Indian Ocean. *J. Geophys. Res. Ocean.* **2020**, *125*, e2019JC015651. [CrossRef]
15. Talley, L.D.; Pickard, G.; Emery, W.; Swift, J. Physical properties of seawater. *Descr. Phys. Oceanogr.* **2011**, *6*, 29–65.
16. Roemmich, D.; Johnson, G.C.; Riser, S.; Davis, R.; Gilson, J.; Owens, W.B.; Garzoli, S.L.; Schmid, C.; Ignaszewski, M. The Argo Program: Observing the global ocean with profiling floats. *Oceanography* **2009**, *22*, 34–43. [CrossRef]
17. Roemmich, D.; Alford, M.H.; Claustre, H.; Johnson, K.; King, B.; Moum, J.; Oke, P.; Owens, W.B.; Pouliquen, S.; Purkey, S.; et al. On the future of Argo: A global, full-depth, multi-disciplinary array. *Front. Mar. Sci.* **2019**, *6*, 439. [CrossRef]
18. Amani, M.; Ghorbanian, A.; Asgarimehr, M.; Yekkehkhany, B.; Moghimi, A.; Jin, S.; Naboureh, A.; Mohseni, F.; Mahdavi, S.; Layegh, N.F. Remote sensing systems for ocean: A review (Part 1: Passive systems). *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 210–234. [CrossRef]
19. Huang, R.X. *Ocean Circulation: Wind-Driven and Thermohaline Processes*; Cambridge University Press: Cambridge, UK, 2010.
20. Munk, W.H. On the wind-driven ocean circulation. *J. Atmos. Sci.* **1950**, *7*, 80–93. [CrossRef]
21. Yan, X.H.; Okubo, A. Three-dimensional analytical model for the mixed layer depth. *J. Geophys. Res. Ocean.* **1992**, *97*, 20201–20226. [CrossRef]
22. Ali, M.; Swain, D.; Weller, R. Estimation of ocean subsurface thermal structure from surface parameters: A neural network approach. *Geophys. Res. Lett.* **2004**, *31*, 021192. [CrossRef]
23. Fu, L.L.; Davidson, R.A. A note on the barotropic response of sea level to time-dependent wind forcing. *J. Geophys. Res. Ocean.* **1995**, *100*, 24955–24963. [CrossRef]
24. Wu, X.; Yan, X.H.; Jo, Y.H.; Liu, W.T. Estimation of subsurface temperature anomaly in the North Atlantic using a self-organizing map neural network. *J. Atmos. Ocean. Technol.* **2012**, *29*, 1675–1688. [CrossRef]
25. Meijers, A.; Bindoff, N.; Rintoul, S. Estimating the four-dimensional structure of the Southern Ocean using satellite altimetry. *J. Atmos. Ocean. Technol.* **2011**, *28*, 548–568. [CrossRef]
26. Guinehut, S.; Dhomps, A.L.; Larnicol, G.; Le Traon, P.Y. High resolution 3-D temperature and salinity fields derived from in situ and satellite observations. *Ocean Sci.* **2012**, *8*, 845–857. [CrossRef]
27. Liu, L.; Peng, S.; Wang, J.; Huang, R.X. Retrieving density and velocity fields of the ocean's interior from surface data. *J. Geophys. Res. Ocean.* **2014**, *119*, 8512–8529. [CrossRef]
28. Wang, J.; Flierl, G.R.; LaCasce, J.H.; McClean, J.L.; Mahadevan, A. Reconstructing the ocean's interior from surface data. *J. Phys. Oceanogr.* **2013**, *43*, 1611–1626. [CrossRef]
29. Akbari, E.; Alavipanah, S.K.; Jeihouni, M.; Hajeb, M.; Haase, D.; Alavipanah, S. A review of ocean/sea subsurface water temperature studies from remote sensing and non-remote sensing methods. *Water* **2017**, *9*, 936. [CrossRef]
30. Holloway, J.; Mengersen, K. Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sens.* **2018**, *10*, 1365. [CrossRef]
31. Jeong, Y.; Hwang, J.; Park, J.; Jang, C.J.; Jo, Y.H. Reconstructed 3-D ocean temperature derived from remotely sensed sea surface measurements for mixed layer depth analysis. *Remote Sens.* **2019**, *11*, 3018. [CrossRef]
32. Maes, C.; Behringer, D.; Reynolds, R.W.; Ji, M. Retrospective analysis of the salinity variability in the western tropical Pacific Ocean using an indirect minimization approach. *J. Atmos. Ocean. Technol.* **2000**, *17*, 512–524. [CrossRef]
33. Nardelli, B.B.; Santoleri, R. Methods for the reconstruction of vertical profiles from surface data: Multivariate analyses, residual GEM, and variable temporal signals in the North Pacific Ocean. *J. Atmos. Ocean. Technol.* **2005**, *22*, 1762–1781. [CrossRef]
34. Su, H.; Huang, L.; Li, W.; Yang, X.; Yan, X.H. Retrieving ocean subsurface temperature using a satellite-based geographically weighted regression model. *J. Geophys. Res. Ocean.* **2018**, *123*, 5180–5193. [CrossRef]
35. Willis, J.K.; Roemmich, D.; Cornuelle, B. Combining altimetric height with broadscale profile data to estimate steric height, heat storage, subsurface temperature, and sea-surface temperature variability. *J. Geophys. Res. Ocean.* **2003**, *108*, JC001755. [CrossRef]
36. Chu, P.C.; Fan, C.; Liu, W.T. Determination of vertical thermal structure from sea surface temperature. *J. Atmos. Ocean. Technol.* **2000**, *17*, 971–979. [CrossRef]
37. Fischer, M. Multivariate projection of ocean surface data onto subsurface sections. *Geophys. Res. Lett.* **2000**, *27*, 755–757. [CrossRef]
38. Lu, W.; Su, H.; Yang, X.; Yan, X.H. Subsurface temperature estimation from remote sensing data using a clustering-neural network method. *Remote Sens. Environ.* **2019**, *229*, 213–222. [CrossRef]
39. Su, H.; Yang, X.; Lu, W.; Yan, X.H. Estimating subsurface thermohaline structure of the global ocean using surface remote sensing observations. *Remote Sens.* **2019**, *11*, 1598. [CrossRef]

40. Su, H.; Zhang, H.; Geng, X.; Qin, T.; Lu, W.; Yan, X.H. OPEN: A new estimation of global ocean heat content for upper 2000 m from remote sensing data. *Remote Sens.* **2020**, *12*, 2294. [CrossRef]
41. Su, H.; Zhang, T.; Lin, M.; Lu, W.; Yan, X.H. Predicting subsurface thermohaline structure from remote sensing data based on long short-term memory neural networks. *Remote Sens. Environ.* **2021**, *260*, 112465. [CrossRef]
42. Song, T.; Wei, W.; Meng, F.; Wang, J.; Han, R.; Xu, D. Inversion of ocean subsurface temperature and salinity fields based on spatio-temporal correlation. *Remote Sens.* **2022**, *14*, 2587. [CrossRef]
43. Xie, H.; Xu, Q.; Cheng, Y.; Yin, X.; Jia, Y. Reconstruction of subsurface temperature field in the south China Sea from satellite observations based on an attention U-net model. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4209319. [CrossRef]
44. Mao, K.; Liu, C.; Zhang, S.; Gao, F. Reconstructing Ocean Subsurface Temperature and Salinity from Sea Surface Information Based on Dual Path Convolutional Neural Networks. *J. Mar. Sci. Eng.* **2023**, *11*, 1030. [CrossRef]
45. Chen, Y.; Liu, L.; Chen, X.; Wei, Z.; Sun, X.; Yuan, C.; Gao, Z. Data driven three-dimensional temperature and salinity anomaly reconstruction of the northwest Pacific Ocean. *Front. Mar. Sci.* **2023**, *10*, 1121334. [CrossRef]
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 03762.
47. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 ×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
48. Huang, B.; Liu, C.; Banzon, V.; Freeman, E.; Graham, G.; Hankins, B.; Smith, T.; Zhang, H.M. Improvements of the daily optimum interpolation sea surface temperature (DOISST) version 2.1. *J. Clim.* **2021**, *34*, 2923–2939. [CrossRef]
49. Kerr, Y.H.; Waldteufel, P.; Wigneron, J.P.; Delwart, S.; Cabot, F.; Boutin, J.; Escorihuela, M.J.; Font, J.; Reul, N.; Gruhier, C.; et al. The SMOS mission: New tool for monitoring key elements ofthe global water cycle. *Proc. IEEE* **2010**, *98*, 666–687. [CrossRef]
50. Hauser, D.; Tourain, C.; Hermozo, L.; Alraddawi, D.; Aouf, L.; Chapron, B.; Dalphinet, A.; Delaye, L.; Dalila, M.; Dormy, E.; et al. New observations from the SWIM radar on-board CFOSAT: Instrument validation and ocean wave measurement assessment. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5–26. [CrossRef]
51. Atlas, R.; Hoffman, R.N.; Ardizzone, J.; Leidner, S.M.; Jusem, J.C.; Smith, D.K.; Gombos, D. A cross-calibrated, multiplatform ocean surface wind velocity product for meteorological and oceanographic applications. *Bull. Am. Meteorol. Soc.* **2011**, *92*, 157–174. [CrossRef]
52. Li, H.; Xu, F.; Zhou, W.; Wang, D.; Wright, J.S.; Liu, Z.; Lin, Y. Development of a global gridded A rgo data set with B arnes successive corrections. *J. Geophys. Res. Ocean.* **2017**, *122*, 866–889. [CrossRef]
53. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 04214.
54. Wang, X.; Jin, Y.; Long, M.; Wang, J.; Jordan, M.I. Transferable normalization: Towards improving transferability of deep neural networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 345446.

**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.