Journal of
Marine Science and Engineering

MDPI

Article

A Method for Sound Speed Profile Prediction Based on CNN-BiLSTM-Attention Network

Zhang Wei, Jin Shaohua *, Bian Gang, Cui Yang, Peng Chengyang and Xia Haixing

Department of Oceanography and Hydrography, Dalian Naval Academy, Dalian 116018, China; 18590361852@163.com (Z.W.); trighosts@163.com (B.G.); 13998435151@163.com (C.Y.); 18908414801@163.com (P.C.); a1466448779@163.com (X.H.)
* Correspondence: jsh_1978@163.com

Abstract: In response to the current challenges in efficiently acquiring sound speed profiles and ensuring their representativeness, considering the need to fully leverage historical sound speed profiles while accounting for their spatiotemporal variability, we introduce a model for sound speed profile prediction based on a CNN-BiLSTM-Attention network, which integrates a convolutional neural network (CNN), a bidirectional long short-term memory network (BiLSTM), and an attention mechanism (AM). The synergy of these components enables the model to extract the spatiotemporal features of sound speed profiles more comprehensively. Utilizing the global ocean Argo grid dataset, the model predicted the sound speed profiles of an experimental zone in the Western Pacific Ocean. In predicting sound speed profiles of a single point, the model achieved a root mean square error (RMSE), relative error (RE), and accuracy (ACC) of 0.72 m/s, 0.029%, and 0.99971, respectively, surpassing comparative models. For regional sound speed profile prediction, the mean RMSE, RE, and ACC of different water layers were 0.919 m/s, −0.016%, and 0.9995, respectively. The experimental outcomes not only confirm the high accuracy of the model, but also highlight its superiority in sound speed profile prediction, particularly as an effective compensatory approach when profile measurements are untimely or contain representational errors.

Keywords: sound speed profile prediction; spatiotemporal features; deep learning; CNN; BiLSTM; attention mechanism

Citation: Wei, Z.; Shaohua, J.; Gang, B.; Yang, C.; Chengyang, P.; Haixing, X. A Method for Sound Speed Profile Prediction Based on CNN-BiLSTM-Attention Network. *J. Mar. Sci. Eng.* **2024**, *12*, 414. https://doi.org/10.3390/jmse12030414

Academic Editor: Marco Cococcioni

Received: 25 January 2024
Revised: 20 February 2024
Accepted: 22 February 2024
Published: 26 February 2024

1. Introduction

Acoustic waves serve as the predominant medium for energy transfer in marine surveying and are extensively utilized across diverse oceanographic domains, including but not limited to underwater acoustic communication, the assessment of marine environmental elements, and the probing of marine resources [1,2]. The propagation of acoustic waves within the oceanic realm is fundamentally determined by the velocity of sound in seawater [3–8]. Due to the intricate nature of the marine environment, the speed at which acoustic waves propagate through a heterogeneous marine medium varies, resulting in acoustic refraction—a crucial factor that impacts the precision of marine exploration endeavors [9–13]. The speed of sound in seawater is influenced by numerous oceanic environmental factors that comprise both relatively stable attributes, such as the depth of the ocean and the constitution of the seabed, and dynamic variables, such as temperature, salinity, and ocean currents [14–16]. These dynamic variables exhibit temporal and spatial fluctuations, leading to significant sound speed variability within seawater over time and space. Therefore, acquiring precise and up-to-date sound speed profiles is a critical objective in marine measurement.

Presently, two methods exist for acquiring sound speed profiles: the direct method and the indirect method [17,18]. The direct method entails utilizing measurement instruments like SVP-plus to directly acquire sound speed profiles at specific locations in seawater. The indirect method employs temperature and salinity devices like CTD, XCTD, and XBT to

acquire temperature and salinity data from seawater. Subsequently, it calculates the sound speed profile using empirical formulas. Owing to the temporal and spatial variability of sound speed and the vastness of the ocean, neither the direct nor the indirect method can efficiently acquire sound speed profiles that accurately represent large areas without significant human and material resources. Consequently, numerous researchers all over the world have suggested utilizing historical sound speed profile data [19] to predict and infer sound speed profiles that closely resemble the actual profiles. This approach seeks to address the limitations of direct measurement methods, enhance operational efficiency, and minimize costs. The research on sound speed profile inversion holds great significance in enhancing the performance of acoustic equipment and fostering the study and development of the ocean [20].

Traditional methods for sound speed profile inversion and prediction typically rely on empirical formulas or statistical analysis models. Reference [21] represents the shallow-water sound speed profile using empirical orthogonal functions (EOFs) and performs sound speed profile inversion by minimizing the error in the arrival angle structure of the signal multipaths. Reference [22] employs the time difference in the arrival of sound rays at a single hydrophone to perform inversion of the shallow-water sound speed profile. Reference [23] utilizes experimental data obtained from the sea to invert sound speed profiles based on the propagation times of sound signals. Reference [24] applies the EOF algorithm to extract the spatial modes of historical sound speed profiles within the measurement area. By combining this with a genetic optimization algorithm and employing terrain distortion as the foundation for constructing a fitness function, the optimal coefficients for reconstruction are obtained, thereby achieving inversion of the sound speed profile. Although traditional methods are capable of performing sound speed profile prediction and inversion, they exhibit certain limitations, such as restricted applicability ranges or insufficient accuracy of the obtained results.

The development of computer technology and the onset of the big data era have led to the application of data-driven machine learning methods for predicting sound speed profiles. This method is capable of adaptive learning and model updating, thus increasing its adaptability. In reference [25], remote sensing parameters were incorporated, and a self-organizing map neural network was combined with a learned dictionary to invert sound speed profiles in the South China Sea. Reference [26] proposed a BP neural network model based on a genetic optimization algorithm to invert sound speed profiles for correcting multi-source depth data. In reference [27], an artificial neural network was utilized to invert sound speed profiles at a point in the Arabian Sea. Reference [28] optimized a BP neural network model using the Levenberg–Marquardt algorithm, and incorporated momentum terms, normalization, and early termination to predict high-precision ocean sound speed profiles, thereby completing the construction of a sound speed field in the South China Sea. Reference [29] proposed a non-linear inversion method based on self-organizing maps, which trained satellite-derived sea surface temperature and height anomaly data in conjunction with EOF coefficients of Argo sound speed profiles to generate mapping graphs, and reconstructed sound speed profiles using the best matching neurons. Machine learning introduces a novel approach for sound speed prediction; however, traditional machine learning necessitates manual feature extraction in the prediction of sound speed profiles. The quality of feature engineering has a direct impact on the accuracy of prediction. Conversely, deep learning places greater emphasis on the model's capability to automatically learn features, thereby effectively addressing this predicament.

The sound speed profile characterizes the changes in sound speed as a function of depth in seawater. The sound speed profile represents a complex multivariate time series that exhibits nonlinearity, non-stationarity, and dynamism, wherein the sound speed at each depth can be regarded as a variable or feature. Consequently, the task of predicting the sound speed profile can be reformulated as a multivariate time series prediction problem. Recurrent neural networks (RNNs) serve as fundamental models for solving time series problems. However, they suffer from the issues of vanishing and exploding gradients,

which result in poor performance for long-term sequence learning [30,31]. Long short-term memory (LSTM), a variant of RNN, overcomes this challenge by incorporating gating mechanisms capable of capturing long-range dependencies and effectively addressing the issue of long-term dependencies in time series data. Bidirectional long short-term memory (BiLSTM) operates on the time series in both forward and backward directions, employing two layers of LSTM networks. It possesses the ability to extract temporal features in a more comprehensive manner compared to unidirectional LSTM. However, both LSTM and BiLSTM have limitations in terms of capturing spatial features. Being a typical spatiotemporal dataset, the sound speed profile cannot be fully captured by a single recurrent neural network or its variations. It frequently necessitates the integration of other networks to attain satisfactory outcomes. Convolutional neural networks (CNNs) have witnessed remarkable achievements in image processing and pattern recognition owing to their capability to capture local patterns and features in data. They have also found applications in time series analysis tasks. Attention mechanism (AM) is a technique that emulates human attention mechanisms, enabling the model to dynamically adapt its focus to various features based on input samples. It proficiently captures the nonlinear relationships among the multivariate time series in the spatiotemporal domain, thereby enhancing the model's comprehension of input data. To fully exploit the benefits offered by CNN, BiLSTM, and AM, and to enhance the prediction accuracy of the sound speed profile, this study proposes a CNN-BiLSTM-Attention network for predicting the sound speed profile, and some literature has also demonstrated the feasibility of this method. Reference [32] proposed a CNN-BiLSTM model to conduct short-term wind power forecasting and achieved greater prediction accuracy. Reference [33] combined a CNN with a BiLSTM to forecast 18 time series of macroeconomic variables of the United States of America. In reference [34], a CNN-BiLSTM-AE model was proposed to improve the accuracy of short-term load forecasting.

The main content arrangement of this article is as follows:

Section 1: Provides an overview of the sound speed profile prediction problem and the need for a combined approach using CNN, BiLSTM, and AM;

Section 2: Introduces the overall architecture of the CNN-BiLSTM-Attention network and provides explanations for each component;

Section 3: Designs the experimental process using the global ocean Argo gridded dataset as the experimental data, and predicts the sound speed profiles for experimental points and areas;

Section 4: Analyzes and evaluates the prediction results of the model, and compares them with the prediction results of other models to demonstrate the effectiveness and superiority of the proposed method;

Section 5: Concludes the research work in this article and discusses possible future improvements and directions for further exploration.

2. CNN-BiLSTM-Attention Model

The model is built on the foundation of the Python programming language. The structure of the CNN-BiLSTM-Attention network is depicted in Figure 1, which comprises several fundamental components, namely, the input layer, the convolutional layer (containing 64 convolutional kernels with a size of 3×3), the pooling layer (employing max pooling with a window size of 2×2), the BiLSTM layer (consisting of two single LSTM layers with 64 units of each), the AM layer, and the output layer. The subsequent section provides an elaborate explanation of each component's structure.

2.1. CNN

CNN, proposed by Yann LeCun et al. in 1998 [35], is a feed-forward neural network with a convolutional structure (Figure 2). It mainly consists of convolutional layers, pooling layers, and fully connected layers [36,37]. The convolutional layer is the core component of CNN, which extracts features from input data using convolutional kernels. The weight

sharing mechanism employed within the convolutional layer not only decreases the computational burden by reducing the number of parameters that require computation, but also enhances the model's ability to capture local features within the input data more effectively. The pooling layer diminishes the dimensionality of the feature maps obtained from the convolutional layer through operations like max pooling or average pooling. As a result, it effectively reduces computational complexity while simultaneously enhancing the model's robustness and generalization capability. The fully connected layer performs global combination and transformation of the extracted features, thereby obtaining higher-level feature representations.

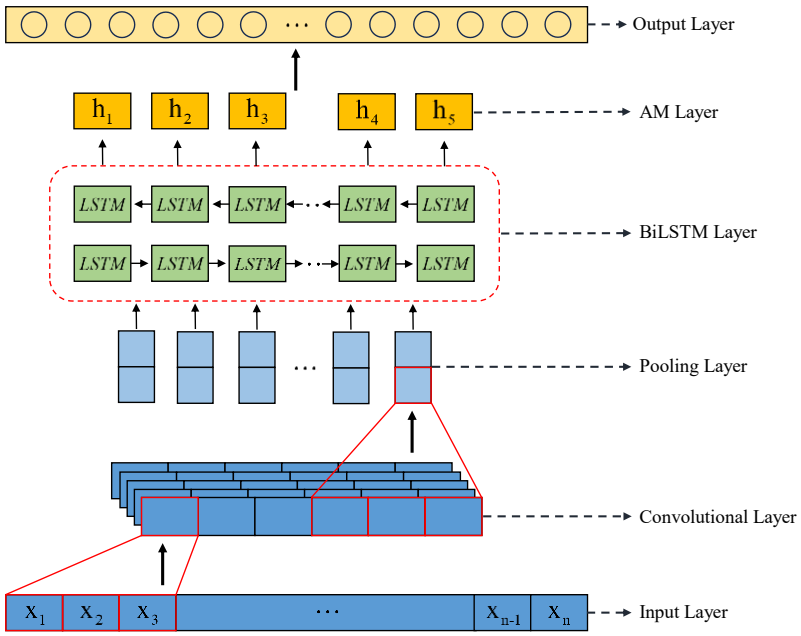

Figure 1. The structure of the CNN-BiLSTM-Attention network.

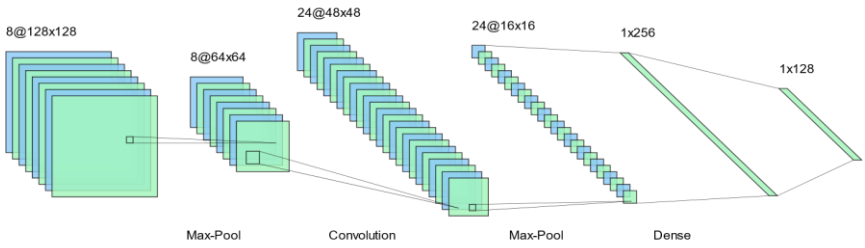

Figure 2. The structure of the CNN network.

2.2. BiLSTM

BiLSTM is developed based on RNN and LSTM. RNN (Figure 3) is a neural network model with memory capability proposed by American computer scientist Rumelhart in 1986 [38]. Compared to other network models, RNN has a unique recurrent structure. At each time step, it simultaneously receives input from the current time step and the output state from the previous time step [39]. This allows it to calculate the output at the current time step and pass the state to the next time step. This recurrent structure empowers RNN to accurately capture the temporal dependencies within the data, which is pivotal for various applications such as natural language processing, machine translation, and speech recognition, among others.

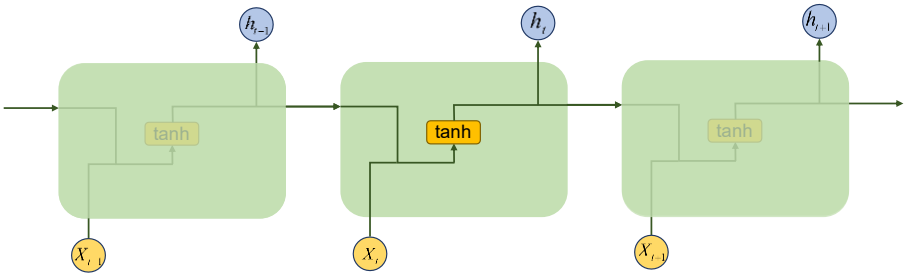

Figure 3. The structure of the RNN network.

One major drawback of RNNs is their susceptibility to the problem of vanishing gradients when processing long sequential data. Gradients in neural networks are mainly utilized to adjust the network weights, and gradient vanishing occurs when the magnitude of gradients decreases during the unfolding over time steps, ultimately tending to zero for earlier time steps. Since RNNs share weights across time steps, the derivatives of the tanh activation function lie between 0 and 1, and if the initial weights are initialized to values less than 1, the product of multiple weight matrices and the tanh function during backpropagation causes the gradients to decay over distance, eventually vanishing and resulting in ineffective training. As a result of this inability to capture long-term dependencies, the performance of the model may be compromised, leading to poor results; consequently, RNNs are characterized as having short-term memory.

To address the problem of vanishing gradients in RNNs, Schmidhuber and others designed the LSTM network in 1997 [40], which is an enhancement of RNNs. The network introduces a gating mechanism to regulate the flow of information, with gate structures autonomously learning to retain significant information and discarding what is unnecessary, thus realizing a more flexible memory mechanism.

Structurally, LSTMs are not drastically different from RNNs, but they employ different functions for computing hidden states. An RNN typically consists of one or more recurrent units connected in a chain-like fashion, with each unit generally containing a simple structure like a tanh layer (Figure 3). LSTMs also have a recursive structure, but their repeating units consist of four interacting layers in a special manner, including three sigmoid layers and one tanh layer (Figure 4). This configuration enables the network to better capture and remember information over long sequences without being affected by issues like vanishing or exploding gradients.

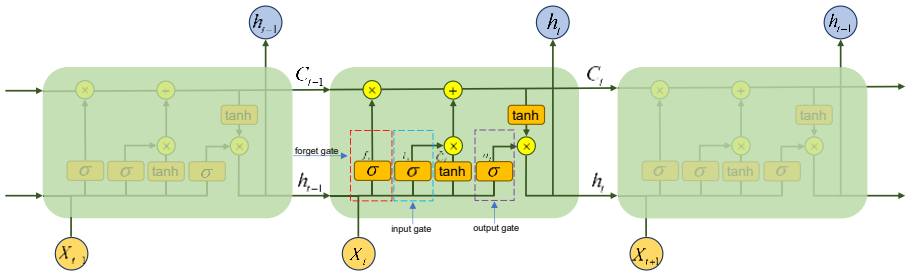

Figure 4. The structure of the LSTM network.

As shown in Figure 4, the LSTM includes three gate units: the forget gate, input gate, and output gate [41]. These three gates, in conjunction with the core of the LSTM—the cell state—collectively control the flow of information. Specifically:

(1) Forget gate. The forget gate is essentially a logic gate that decides which information should be discarded from the cell state by applying a sigmoid function to the input of the current time step and the hidden state from the previous time step. The formula is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

Here, f_t is the output of the forget gate, with values ranging from 0 to 1. W_f represents the weights of the forget gate, h_{t-1} is the hidden state from the previous time step, x_t is the input at the current time step, and b_f is the bias term of the forget gate.

(2) Input gate. The input gate decides which information will be stored in the cell state. The sigmoid function determines which information is to be updated, while the tanh function calculates the output of the candidate input gate. The formula is as follows:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \widetilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{aligned} \tag{2}$$

Here, i_t is the output of the input gate with values ranging from 0 to 1, W_i represents the weights of the input gate, b_i is the bias term of the input gate, $\widetilde{C}_t$ is the output of the candidate input gate, W_C represents the weights of the candidate input gate, and b_C is the bias term of the candidate input gate.

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{3}$$

(3) Cell state update. The outputs of the forget gate, input gate, and candidate input gate are combined with the cell state of the previous time step to update the cell state for the current time step. The formula is as follows:

Here, C_t is the updated cell state for the current time step, and C_{t-1} is the cell state from the previous time step.

(4) Output gate. The final output is determined by considering the input of the current time step, the hidden state from the previous time step, and the updated cell state. The formula is as follows:

Here, o_t is the output of the output gate with values between 0 to 1, W_o represents the weights of the output gate, b_o is the bias term of the output gate, and h_t is the current time step's hidden state.

The concept of BiLSTM was initially proposed by Schuster in 1997 [42]. This architecture comprises two separate LSTMs (Figure 5): one processes the sequence in the forward direction, while the other handles it in the backward direction. Subsequently, the outputs of the two LSTMs are concatenated, enabling a more comprehensive grasp of the contextual information within the sequence and endowing the model with enhanced representational capacity.

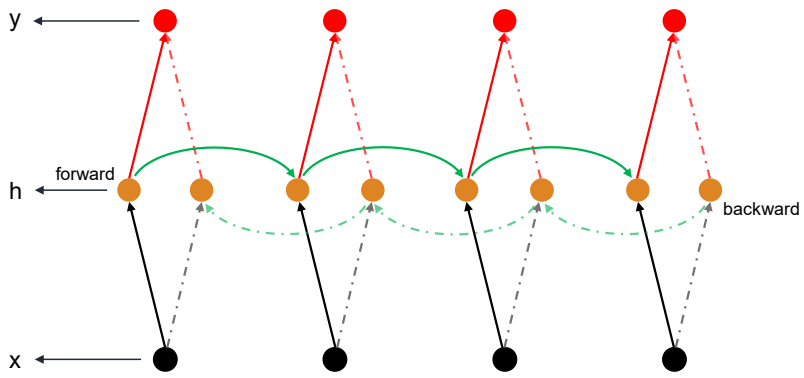

Figure 5. The structure of the BiLSTM network.

2.3. AM

The concept of attention mechanisms in deep learning essentially resembles human selective visual attention, with the core objective of extracting relevant important information from a large pool of data. Vaswani et al. define attention as a method that maps a query and a set of key–value pairs to an output [43]. The computation process of the attention mechanism typically includes three stages, as shown in Figure 6.

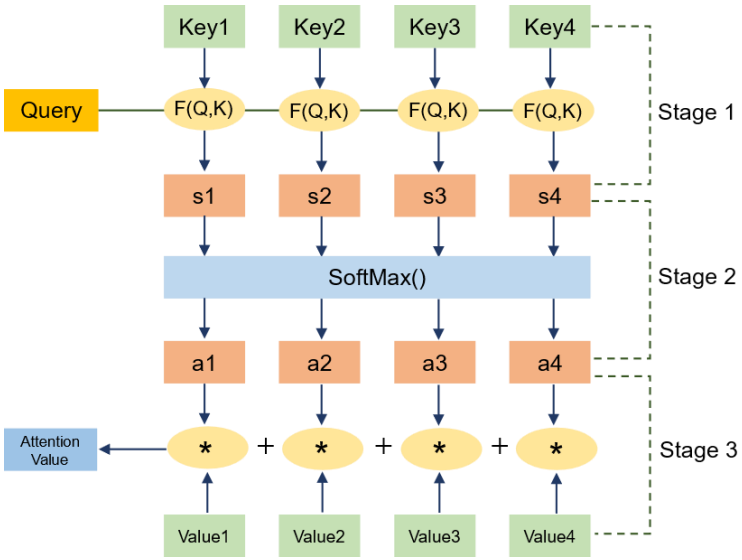

Figure 6. Attention mechanism.

Taking the commonly used dot-product attention as an example, the calculation process is as follows:

(1) The first stage: Utilize the dot product of vectors to compute the similarity between the Query and Key, also known as the attention score;

(2) The second stage: Use the Softmax function to normalize the attention scores, obtaining the attention weights;

(3) The third stage: Weighted summation of the Key's weights and their corresponding Values to output the final result. The specific calculation formula is as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \tag{4}$$

Here, Q represents the query vector, K is the key vector, V is the value vector, and d_K is the dimension of the key vector.

3. Materials and Experiments

3.1. Data Source

The data used in this experiment come from the global ocean Argo gridded dataset (BOA-Argo), released by the Hangzhou Global Ocean Argo System Field Observation Research Station under the auspices of the Ministry of Natural Resources of China (http://www.argo.org.cn, accessed on 31 October 2023). The research personnel at the station have meticulously performed rigorous quality re-control on the Argo profile data utilized in this dataset, guaranteeing the utmost quality of the original data.

Currently, this dataset provides a total of 234 monthly gridded profile data on global ocean (longitude: 180° W–180° E, latitude: 80° S–80° N) temperature and salinity from January 2004 to June 2023. The horizontal resolution is 1° × 1° (longitude: 0.5:1.0:359.5, latitude: −79.5:1.0:79.5), while the vertical direction consists of 58 standard layers with a maximum pressure of 1975 dbar. The data are stored on a monthly basis, where each month has an individual file dedicated to it. The dataset can be downloaded for free in both MATLAB and NetCDF formats.

3.2. Experimental Data

In order to validate the effectiveness of the model in the spatiotemporal prediction of sound speed profiles, we selected an experimental area in the Western Pacific Ocean (126.5° E–135.5° E, 13.5° N–22.5° N) and a specific point outside this region (122.5° E, 22.5°

N), as shown in Figure 7. The historical dataset covers a time period from January 2004 to December 2022, which corresponds to 19 years and a total of 228 months.

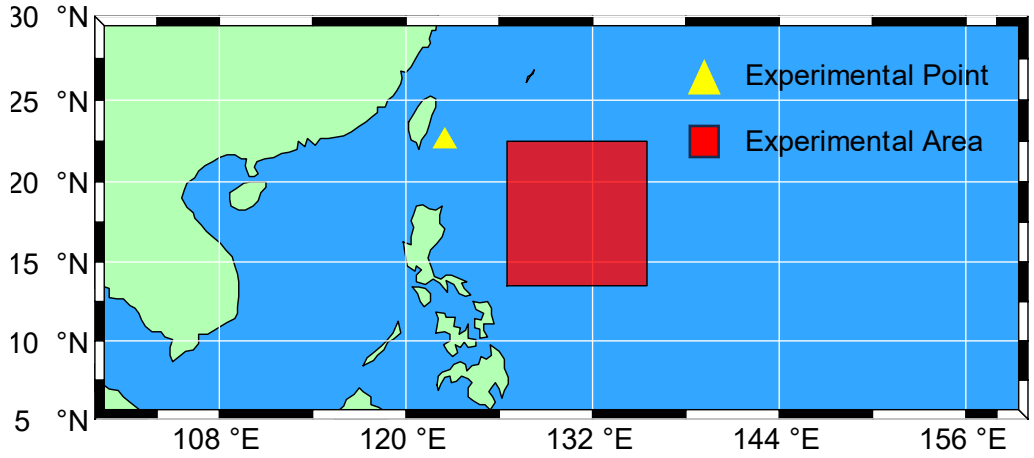

Figure 7. The experimental point and area.

Given that the BOA-Argo dataset only includes temperature, salinity, and depth data, it is crucial to choose a suitable empirical formula for sound speed calculation in order to obtain the sound speed profiles of the experimental object. Reference [44] compared the adaptability of seven empirical sound speed formulas, and it was concluded that Mackenzie's formula ensures accurate computational results across all depth levels, making it an ideal choice for calculating full-depth sound speed profiles. Additionally, Mackenzie's formula is relatively simple, making it the chosen method for sound speed calculation in this paper. The formula is as follows:

$$C = 1448.96 + 4.591T - 5.304 \times 10^{-2}T^2 + 2.374 \times 10^{-4}T^3 + 1.340 \times (S-35) + 1.630 \times 10^{-2}D + 1.675 \times 10^{-7}D^2 - 1.025 \times 10^{-2}T(S-35) - 7.139 \times 10^{-13}TD^3 \quad (5)$$

Here, C (m/s) represents sound speed, T (°C) is temperature, S(ppt) is salinity, and D (m) is depth.

The historical sound speed profile data for the experimental point and area were computed, and the statistical parameters are presented in Table 1.

Table 1. Statistical parameters of historical sound speed profiles for the experimental point and experimental area.

Experimental Object	Maximum Sound Speed (m/s)	Minimum Sound Speed (m/s)	Average Value (m/s)	Data Dimension
Experimental Point	1546.67	1480.30	1508.72	228 × 58
Experimental Area	1548.00	1479.81	1508.85	228 × 10 × 10 × 58

Monthly average sound speed profiles for the experimental point and area are shown in Figure 8.

3.3. Experimental Process

The experimental design process is illustrated in Figure 9 and can be divided into three main parts: dataset splitting, model training, and model prediction. The following section will provide a brief overview of each part.

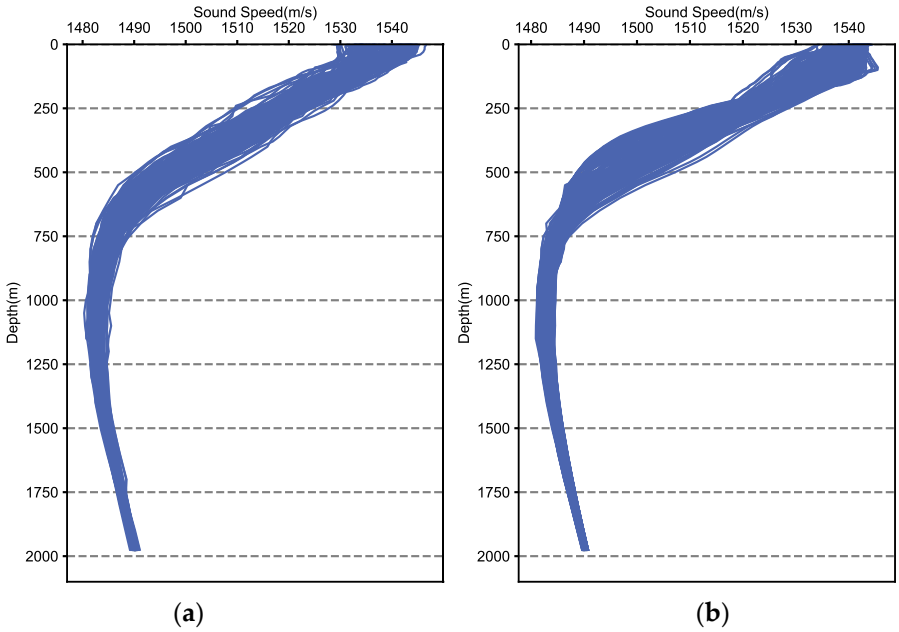

Figure 8. (**a**) Monthly average sound speed profiles for the experimental point; (**b**) monthly average sound speed profiles for the experimental area.

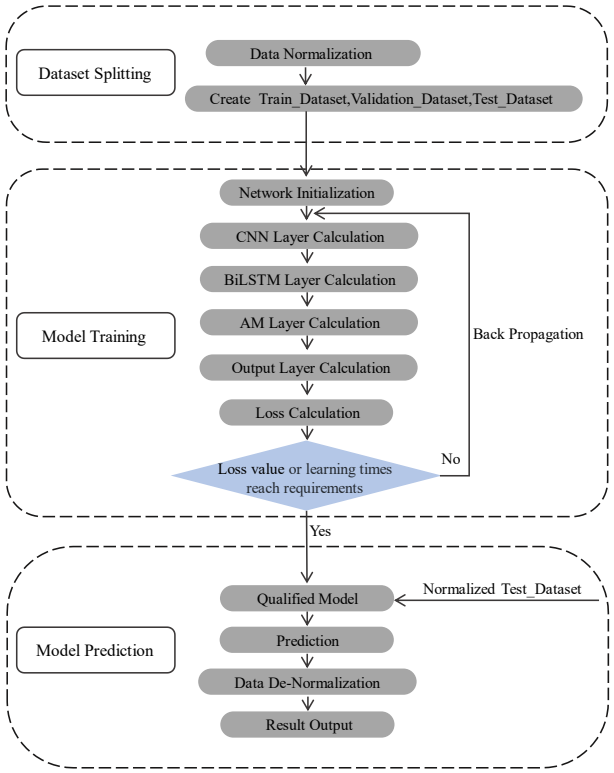

Figure 9. Experimental design process.

3.3.1. Dataset Splitting

The time step is a crucial hyperparameter in time series forecasting models, referring to the number of historical sample points used to predict future data. Selecting an appropriate time step is essential for the accuracy of time series prediction. The optimal time step is usually determined through iterative experiments, taking into account the characteristics of the data and the forecasting requirements. Previous research [45] has explored different time

steps of 1, 6, 24, and 28 months in sound speed profile prediction experiments. The findings revealed that employing a time step of 24 months produced the most favorable outcomes for predicting sound speed profiles, exhibiting exceptional performance across multiple evaluation metrics. The predicted sound speed profiles exhibited a close resemblance to the actual profiles. Based on this finding, this study adopts a time step of 24 for dataset splitting.

As depicted in Figure 10, the sound speed profile data were partitioned into groups of 24 consecutive months as inputs, with the subsequent month's data serving as the output, resulting in the formation of the training dataset. In the course of model training, 20% of the data was randomly chosen from the training dataset as the validation dataset. As for the test dataset, the last 25 months of sound speed profile data were employed, with the first 24 months serving as inputs and the last month as the output. The dimensions of each dataset are shown in Table 2.

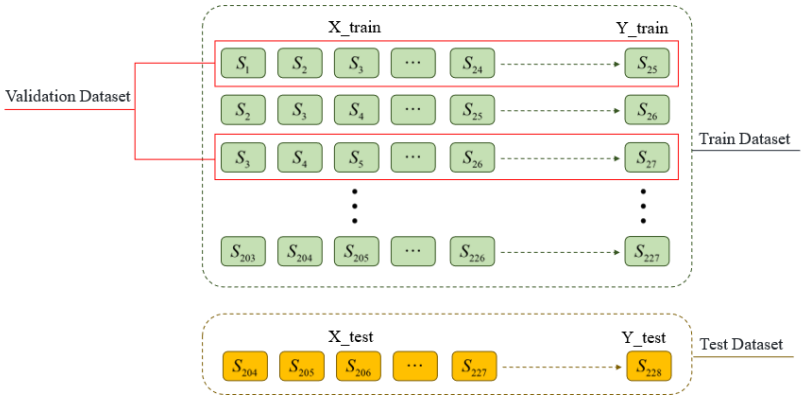

Figure 10. Splitting of the training dataset, validation dataset, and test dataset.

Table 2. Dimensions of each dataset.

Experimental Object	Experimental Point	Experimental Area
Input of training and validation dataset	203 × 24 × 58	203 × 24 × 10 × 10 × 58
Output of training and validation dataset	203 × 58	203 × 10 × 10 × 58
Input of test dataset	1 × 24 × 58	1 × 24 × 10 × 10 × 58
Output of test dataset	1 × 58	1 × 10 × 10 × 58

3.3.2. Model Training

The model training process consisted of the following steps:

(1) Data normalization: Normalization is a common data preprocessing method in machine learning. Since the original sound speed data have large values, they are not conducive to model training. Normalization can scale the data to a specified range and accelerate model convergence. In this study, the Min–Max normalization method is applied to the whole dataset to map the data to the range of [0, 1]. The calculation formula is as follows:

$$X' = (X - min)/(max - min) \quad (6)$$

Here, X' is the normalized data, X is the original data, max is the maximum value of the original data, and min is the minimum value of the original data.

(2) Network weight initialization: Effective weight initialization prior to model training can accelerate convergence and enhance model performance. In this study, we employed the Glorot uniform initialization method, which effectively preserves signal stability during both the forward and backward propagation processes.

(3) Feature extraction: After passing through the input layer, the data undergo a series of operations in the CNN layer, encompassing convolution and pooling, to extract local spatial features and generate feature maps. Subsequently, these feature maps are forwarded to the BiLSTM layer, responsible for analyzing long-range dependencies within

the data and transferring the captured temporal features to the AM layer. The AM layer seizes upon the significance of distinct time steps for the model output. Particularly when confronted with extensive time series, it aids the model in focusing on pivotal components and mitigating information loss. The features extracted from each layer are then fed into the fully connected layer (output layer), which yields the model's prediction data.

(4) Loss computation: The computation of loss involves evaluating the discrepancy between the predicted and true values, employing a pre-defined loss function. For this study, the mean squared error (MSE) is employed as the loss function. Consequently, the network weights and biases are iteratively adjusted through the backpropagation process.

(5) Model output: The training process ceases when either the maximum specified number of iterations is attained or when the loss descends below a predetermined threshold, yielding the final network parameters.

Model Prediction

The process of model prediction entails feeding the normalized test dataset into the trained model. The model generates predictions, which are subsequently transformed through denormalization to derive the ultimate predicted sound speed profiles. The denormalization formula is as follows:

$$X = X' \times (max - min) + min \tag{7}$$

Here, X is the denormalized data, X' is the model's output data, max is the maximum value of the output data, and min is the minimum value of the output data.

3.4. Model Evaluation

To comprehensively and objectively assess the predictive performance of various models, this study embraces the utilization of evaluation metrics such as the root mean squared error (RMSE), relative error (RE), and accuracy (ACC) to evaluate the model's quality.

RMSE measures the average error between the predicted values and the true values. A smaller RMSE indicates a lower prediction error and a better model performance.

RE reflects the relative difference between the predicted values and the true values. It is usually expressed as a percentage, and a smaller absolute value indicates a smaller deviation of the predicted values from the true values, indicating a better prediction performance of the model.

ACC measures the accuracy of model predictions. A value closer to 1 indicates higher prediction accuracy.

The formulas for calculating these metrics are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{8}$$

$$RE = \frac{\hat{y}_i - y_i}{y_i} \times 100\% \tag{9}$$

$$ACC = 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|/y_i}{n} \tag{10}$$

Here, y_i is the true value and $\hat{y}_i$ is the predicted value.

4. Results and Discussion

In order to validate the effectiveness and superiority of the proposed method in this study, we conducted synchronous single-point sound speed profile predictions using the CNN, LSTM, CNN-LSTM, CNN-BiLSTM, and CNN-LSTM-Attention models. Here, the CNN model contained 64 convolutional kernels with a size of 3. The LSTM model contained 64 units. The CNN-LSTM and CNN-LSTM-Attention model both consisted of a CNN layer

and an LSTM layer. The CNN layer included 64 convolutional kernels of size 3, while the LSTM layer consisted of 64 units. The CNN-BiLSTM model included a CNN layer and a BiLSTM layer. The CNN layer included 64 convolutional kernels of size 3, while the BiLSTM layer consisted of two single LSTM layers with 64 units of each.

We compared the prediction results obtained from different models, and subsequently employed the CNN-BiLSTM-Attention model independently to forecast the sound speed within the experimental area. The analysis of the experimental results is presented below.

4.1. Analysis of Sound Speed Profile Prediction Results for the Experimental Point

Figure 11 depicts a comparison of the prediction results obtained from the CNN, LSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention, and CNN-BiLSTM-Attention models with the actual sound speed profiles. Visually, the prediction results of all models aligned with the trends observed in the actual profiles. The prediction results obtained from the CNN-BiLSTM-Attention model exhibited the highest level of agreement with the actual values, whereas the CNN model exhibited the largest discrepancy. Specifically, all models demonstrated strong performance in terms of predicting sound speed values within the deep isothermal layer (the layer characterized by the lowest sound speed values and the water bodies beneath it), given the relatively stable temperature of these water bodies and the dominant influence of depth on sound speed. The relationship between sound speed and depth in this layer was relatively straightforward, exhibiting a linear change with depth, and all models were capable of capturing this characteristic. The prediction results of the models exhibited significant variations in the thermocline layer (approximately 50–1000 m deep), where sound speed is primarily influenced by temperature. The variation in this layer was more complex, resulting in a sound speed profile that visually appeared as a curve with intricate changes. The different mapping abilities of the models for this relationship led to variations in the prediction results.

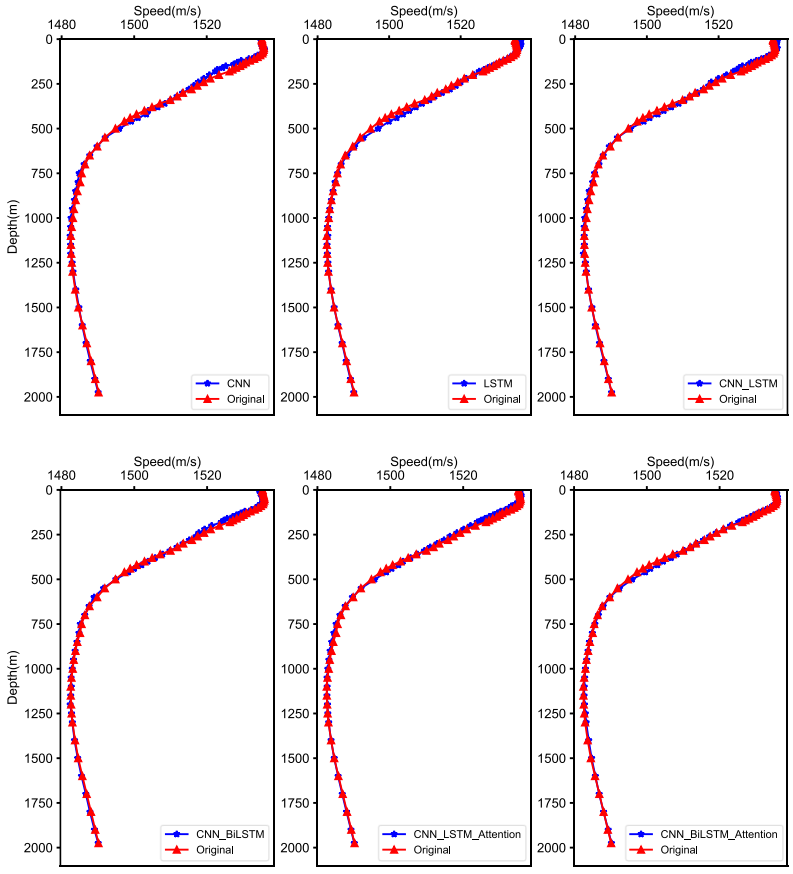

Figure 11. Comparison between predicted results of different models and actual values.

The RMSE, RE, and ACC of the prediction results for each model are shown in Table 3. The RMSE values for the CNN, LSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention, and CNN-BiLSTM ranged from 1.46 m/s to 0.72 m/s. The RMSE values for the CNN, LSTM, CNN-LSTM, CNN-BiLSTM, CNN-LSTM-Attention, and CNN-BiLSTM ranged from 1.46 m/s to 0.72 m/s. This indicates that the ranking of model performance is not uniquely determined by a single evaluation metric. Overall, irrespective of the evaluation metric used, the CNN-BiLSTM-Attention model outperformed the other models, whereas the CNN model had evident limitations. These results provide data support for the previous qualitative assessments.

Table 3. RMSE, RE, and ACC of each model's prediction results.

Experimental Object	RMSE (m/s)	RE (%)	ACC
CNN	1.46	−0.061	0.99939
LSTM	1.14	0.048	0.99952
CNN-LSTM	1.06	−0.050	0.99950
CNN-BiLSTM	0.99	−0.046	0.99954
CNN-LSTM-Attention	0.97	−0.046	0.99954
CNN-BiLSTM-Attention	0.72	0.029	0.99971

4.2. Analysis of Sound Speed Profile Prediction Results in the Experimental Area

By conducting analyses from both horizontal and vertical perspectives, our objective was to further showcase the superiority of the proposed method in terms of capturing the spatiotemporal characteristics of sound speed profiles.

4.2.1. Analysis of Prediction Results for Different Water Layers

The RMSE, RE, and ACC for different water layers were calculated using formulas 8, 9, and 10, respectively, as depicted in Figure 12.

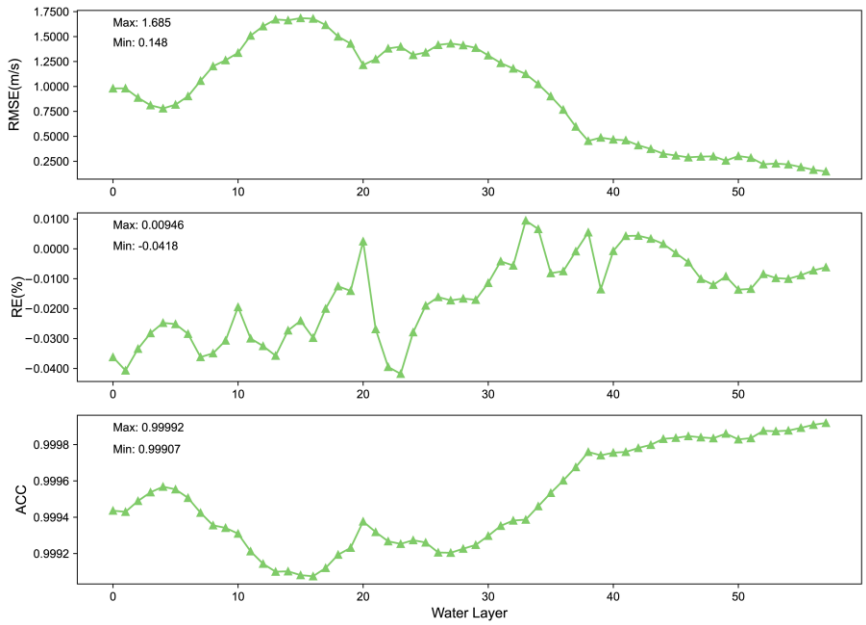

Figure 12. RMSE, RE, and ACC of different water layers.

Table 4 presents the maximum, minimum, and mean values of each metric. We can see in Figure 12 that in general, the accuracy of the model's sound speed predictions for diverse water layers exhibited a pattern of an initial increase, a subsequent decrease, and an eventual increase with increasing water depth. Among all water layers, the predictions for the final layer demonstrated the highest accuracy, suggesting the model's proficiency

in forecasting sound speed in deep water bodies (approximately 800 m or deeper). Conversely, shallow water bodies manifested intricate fluctuations in sound speed, owing to intricate alterations in oceanic environmental factors. Consequently, the model exhibited an unsatisfactory performance in this section, with the lowest prediction accuracy observed in the 10th to 20th layers (approximately 90−200 m deep).

Table 4. Maximum, minimum, and mean values of each metric.

Metric	Max	Min	Mean
RMSE (m/s)	1.685	0.148	0.919
RE (%)	0.00946	−0.0418	−0.016
ACC	0.99992	0.99907	0.9995

Figure 13 illustrates the prediction errors of sound speed for each point in the four water layers at depths of 0 m, 140 m, 500 m, and 1000 m. The selection of 140 m was based on its highest RMSE and unsatisfactory predictive performance, rendering it a representative depth. From the figure, it can be seen that while the prediction errors varied among different layers, the regions with higher prediction errors primarily occurred at the boundaries of the studied area, suggesting that the model exhibited superior performance in predicting sound speed within the region as compared to the edges.

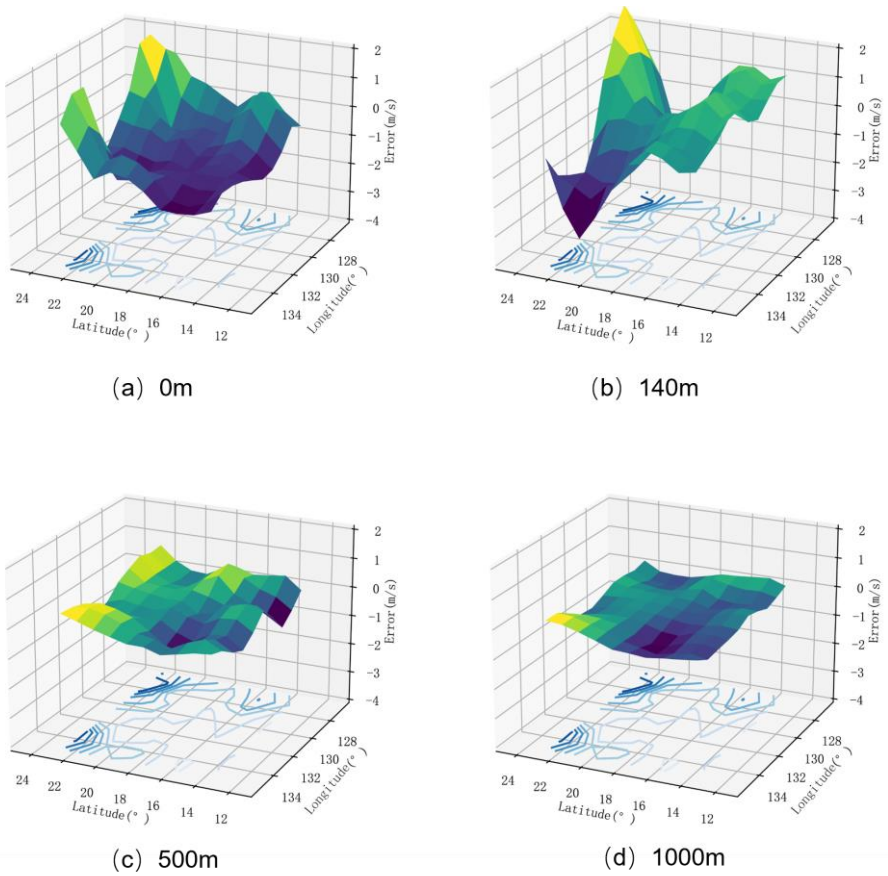

Figure 13. Prediction errors of sound speed for each point in the four water layers at depths of 0 m (**a**), 140 m (**b**), 500 m (**c**), and 1000 m (**d**).

4.2.2. Analysis of Prediction Results for Different Sound Speed Profiles

The model predicted sound speed profiles for a total of 100 location points, and the RMSE, RE, and ACC were calculated for each sound speed profile. The results are shown in Figure 14, and the maximum, minimum, and mean values of each metric are

presented in Table 5. Observing the figure, it is evident that, apart from a few edge points showing significant prediction errors in the sound speed profiles, the variations in metrics among the predicted sound speed profiles at other locations were minimal and statistically indistinguishable. In conjunction with Section 4.2.1, this finding reaffirms the superior accuracy of the model in predicting sound speed within the region compared to the edges, thereby indicating the necessity of enhancing the model's capability to capture sound speed features at edge positions. Furthermore, the average value of RE amounted to −0.016%, implying that the predicted sound speed profiles tended to underestimate the actual values.

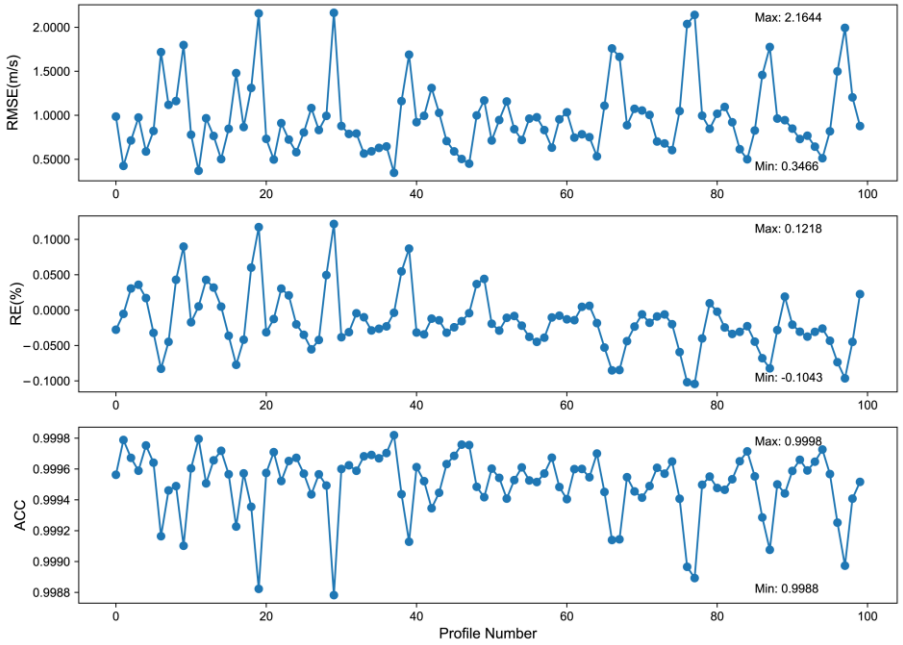

Figure 14. RMSE, RE, and ACC of different sound speed profiles.

Table 5. Maximum, minimum, and mean values of each metric.

Metric	Max	Min	Mean
RMSE (m/s)	2.164	0.3466	0.966
RE (%)	0.1218	−0.1043	−0.016
ACC	0.99992	0.99907	0.9995

Four random position points were randomly selected from the experimental area, and their corresponding coordinates are presented in Table 6. Figure 15 illustrates the comparison between the predicted sound speed profiles and the original sound speed profiles at these selected points. From the figure, it can be observed that the predicted sound speed profiles for points 1 and 2 closely matched the original sound speed profile. However, points 0 and 3 displayed substantial deviations in the predicted sound speed profiles within the depth range of 200 to 500 m. Considering the location coordinates and referring to Figure 12, it can be inferred that points 1 and 2 were situated within the region where the model exhibited minimal prediction errors. Conversely, points 0 and 3 were in proximity to the experimental area's boundary and lay within the region characterized by significant variations in the model's prediction errors.

The predicted errors for each sound speed point in the four profiles are visualized in a 3D graph, depicted in Figure 16. The color bar in the graph corresponds to the predicted errors for each sound speed point. The graph reveals that the error distribution among the sound speed points in profiles 1 and 2 was relatively consistent, primarily ranging from −1 to 1 m/s. The positive and negative errors exhibited a symmetrical distribution. On the other hand, profiles 0 and 3 had predominantly negative errors. The sound speed

points with larger errors were mainly distributed in the 20th to 35th layers (approximately 200 to 550 m in depth). The maximum error in profile 0 occurred at the 27th layer, with a magnitude of −2.13 m/s, while the maximum error in profile 3 occurred at the 26th layer, with a magnitude of −1.57 m/s.

Table 6. Positions of selected sound speed profiles.

Position Point Id	Longitude	Latitude
0	126.5	18.5
1	129.5	16.5
2	129.5	20.5
3	128.5	18.5

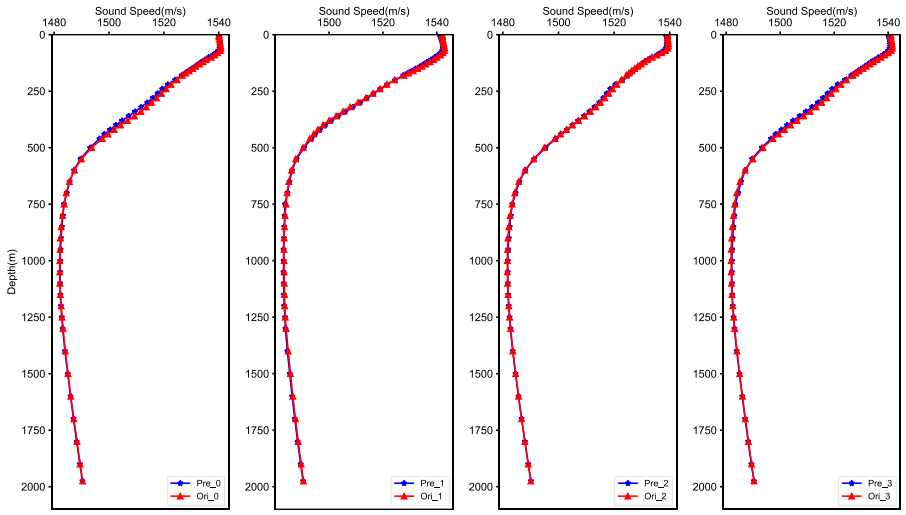

Figure 15. Comparison between predicted results of different positions and actual values.

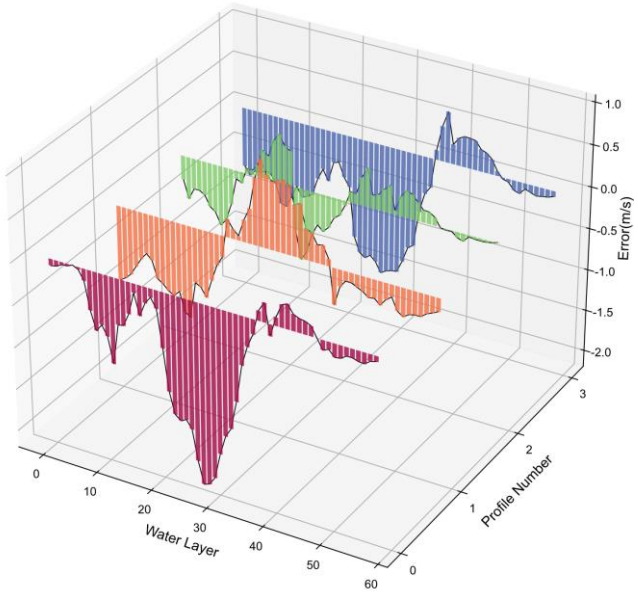

Figure 16. Predicted errors for each sound speed point in the four sound speed profiles.

5. Conclusions

Prompt and precise acquisition of sound speed profiles in designated sea areas is crucial for improving the accuracy of marine measurement equipment. Currently, the existing methods for acquiring sound speed profiles exhibit deficiencies in terms of obtaining

comprehensive and representative data. To tackle the challenges associated with delayed measurements, complexities in data collection, and the assurance of representative sound speed data, this study proposes a sound speed profile prediction method based on the CNN-BiLSTM-Attention network. This integrated network combines convolutional neural networks, bidirectional long short-term memory networks, and attention mechanisms to effectively extract the spatiotemporal features of sound speed profiles, with a focus on the influences of different time steps on the prediction results.

In this study, the global ocean Argo gridded dataset was used for the experimental data. A specific area in the Western Pacific Ocean, along with a point outside the region, was chosen as the experimental object. A total of six individual or combined models, including the CNN-BiLSTM-Attention model, were used to predict the sound speed profiles simultaneously at the experimental points. The results showed that the proposed model achieved a better performance compared to other models, with RMSE, RE, and accuracy (ACC) values of 0.72 m/s, 0.029%, and 0.99971, respectively. These findings demonstrate both the accuracy and superiority of the proposed model, as well as the effectiveness of employing a combined model to enhance prediction accuracy compared to utilizing a single model.

When solely employing the CNN-BiLSTM-Attention model to predict sound speed profiles within the experimental area, the average values of RMSE, RE, and ACC for different water layers were 0.919 m/s, −0.016%, and 0.9995, respectively. The average values of RMSE, RE, and ACC for different sound speed profiles were recorded as 0.966 m/s, −0.016%, and 0.9995, respectively. The experimental findings indicated that the model exhibited superior performance in predicting sound speed in deeper water layers compared to shallower ones, and the prediction accuracy was higher for sound speed within the central region of the area in comparison to its peripheral edges. Consequently, it is recommended to broaden the spatial extent of historical data when utilizing this model to forecast sound speed profiles within a specific area, as doing so can enhance the accuracy of the predictions.

The research findings of this study offer a potential remedy for promptly acquiring sound speed profiles or addressing representative errors that may arise during the actual measurement process, thereby offering practical significance. In the future, further optimization of the model will be conducted to improve the accuracy of sound speed prediction in shallow water areas.

Author Contributions: Conceptualization: J.S.; writing and methodology: Z.W.; formal analysis: B.G. and C.Y.; data collection: P.C. and X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Heidemann, J.; Stojanovic, M.; Zorzi, M. Underwater sensor networks: Applications, advances and challenges. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2012**, *370*, 158–175. [CrossRef] [PubMed]
2. Ahmed, A.; Younis, M. Distributed real-time sound speed profiling in underwater environments. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–7.
3. Zhang, X.; Zhang, Y.; Zhang, J.; Nie, B.S.; Yao, Z.S. EOF Analysis of Sound Speed Profile Sequences in the Waters East of Taiwan. *Adv. Mar. Sci.* **2010**, *28*, 498–506.
4. Li, Q. Progress in Acoustic Research. *J. Acoust.* **2001**, *4*, 295–301.
5. Li, J. *Principles, Techniques and Methods of Multibeam Surveying*; Ocean Press: Beijing, China, 1999.

6. He, G.; Liu, F.L.; Yu, P.; Yang, S.X.; Zhang, Z.R.; Zhao, Z.B. Sound Speed Correction in Multibeam Echosounder Systems. *Mar. Geol. Quat. Geol.* **2000**, *4*, 109–114.
7. Zhao, J.; Liu, J. *Multibeam Bathymetric and Image Data Processing*; Wuhan University Press: Wuhan, China, 2008.
8. Zhou, J.; Zhou, Q.; Lü, L.; Chen, C.; Huang, H.L. Discussion on the Correction of Sound Speed Profiles in Multibeam Surveys. *Hydrogr. Surv. Charting* **2014**, *34*, 62–65+68.
9. Ding, J.; Zhou, X.; Tang, Q. Correction Technique for Ray Refraction of Multibeam Echosounder Systems Based on Equivalent Sound Speed Profile Method. *Hydrogr. Surv. Charting* **2004**, *6*, 27–29.
10. Capell, W.J. Determination of sound velocity profile errors using multibeam data. In Proceedings of the Oceans '99. MTS/IEEE. Riding the Crest into the 21st Century, Conference and Exhibition, Conference Proceedings (IEEE Cat. No.99CH37008), Seattle, WA, USA, 13–16 September 1999; IEEE & Marine Technol. Soc: Piscataway, NJ, USA, 1999; Volume 3, pp. 1144–1148.
11. Tonchia, H.; Bisquay, H. The effect of sound velocity on wide swath multibeam system data. In Proceedings of the OCEANS 96 MTS/IEEE Conference Proceedings, The Coastal Ocean—Prospects for the 21st Century, Fort Lauderdale, FL, USA, 23–26 September 1996; IEEE: Piscataway, NJ, USA, 1996; Volume 2, pp. 969–974.
12. Dinn, D.F.; Loncarevic, B.D.; Costello, G. The effect of sound velocity errors on multi-beam sonar depth accuracy. In Proceedings of the "Challenges of Our Changing Global Environment", Conference Proceedings, OCEANS '95 MTS/IEEE, San Diego, CA, USA, 9–12 October 1995; IEEE: Piscataway, NJ, USA, 1995; Volume 2, pp. 1001–1010.
13. Park, J.C.; Kennedy, R.M. Remote sensing of ocean sound speed profiles by a perceptron neural network. *IEEE J. Ocean. Eng.* **1996**, *21*, 216–224. [CrossRef]
14. Wang, T.; Su, L.; Ren, Q.; Wang, W. Application of Recurrent Neural Network in the Joint Inversion of Shallow Water Sound Speed and Sound Source. *J. Harbin Eng. Univ.* **2021**, *42*, 1133–1139.
15. Hu, J. *Sound Speed Profile Inversion Based on RBF Neural Network and Software Implementation*; Xiangtan University: Xiangtan, China, 2018.
16. Wang, T.; Su, L.; Ren, Q.; Wang, W. Prediction Method for Full Ocean Depth Sound Speed Profiles Based on Attention Mechanism. *J. Electron. Inf. Technol.* **2022**, *44*, 726–736.
17. Huang, C.; Lu, X.; Ye, A.; Luo, S. Quality Control and Evaluation of Seafloor Topographic Measurement Results (II): Acquisition of Sound Speed Profiles in Deep and Distant Sea Areas. *Hydrogr. Surv. Charting* **2017**, *37*, 12–16+20.
18. Zhang, Q.; Chen, X.; Liu, Q. Research on Sound Speed Profile Acquisition Methods in Offshore Multibeam Bathymetric Survey. *Hydrogr. Surv. Charting* **2019**, *39*, 1–4.
19. Carnes, M.R.; Mitchell, J.L.; De Witt, P.W. Synthetic temperature profiles derived from Geosat altimetry: Comparison with air-dropped expendable bathythermograph profiles. *J. Geophys. Res. Ocean.* **1990**, *95*, 17979–17992. [CrossRef]
20. Sun, J.; Zhang, J.; Tang, Y. Sound Speed Profile Inversion Method Using Double Population Constrained QPSO-BP. *Sci. Surv. Mapp.* **2021**, *46*, 127–134.
21. Shen, Y.; Ma, Y.; Tu, Q.; Jiang, X. Inversion Method and Experimental Verification of Shallow Water Sound Speed Profile. *J. Northwestern Polytech. Univ.* **2000**, *2*, 212–215.
22. Zhang, Z.; Ma, Y.; Ni, J.; Tong, L. Inversion of Shallow Water Sound Speed Profile Based on the Time Difference of Acoustic Ray Arrival. *J. Northwestern Polytech. Univ.* **2002**, *1*, 36–39.
23. Tang, J.; Yang, S. Inversion of Sound Speed Profile in Seawater by Propagation Time. *J. Harbin Eng. Univ.* **2006**, *5*, 733–736+756.
24. Sun, W.; Bao, J.; Jin, S.; Xiao, F. Correction of Multibeam Seafloor Topographic Distortion and Sound Speed Profile Inversion. *Geomat. Inf. Sci. Wuhan Univ.* **2016**, *41*, 349–355.
25. Xie, L.; Liu, C.; Liang, W. Reconstruction and Inversion of Sound Speed Profiles Based on Dictionary Learning. *J. Mar. Technol.* **2023**, *42*, 12–19.
26. Yuan, H.; Jia, S.; Jin, S.; Zhang, L.; Wang, H. Sound Speed Correction of Crowd-Sourced Bathymetric Data Using GA-NN Model Inversion of Sound Speed Profiles. *Geomat. Inf. Sci. Wuhan Univ.* **2023**, *48*, 377–385. [CrossRef]
27. Jain, S.; Ali, M.M. Estimation of Sound Speed Profiles Using Artificial Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2006**, *3*, 467–470. [CrossRef]
28. Huang, J.; Luo, Y.; Shi, J.; Ma, X.; Li, Q.Q.; Li, Y.Y. Rapid Modeling of the Sound Speed Field in the South China Sea Based on a Comprehensive Optimal LM-BP Artificial Neural Network. *J. Mar. Sci. Eng.* **2021**, *9*, 488. [CrossRef]
29. Li, H.; Qu, K.; Zhou, J. Reconstructing Sound Speed Profile from Remote Sensing Data: Nonlinear Inversion Based on Self-Organizing Map. *IEEE Access* **2021**, *9*, 109754–109762. [CrossRef]
30. Cao, Y.; Xie, M. Stock Price Prediction Analysis Based on WD-CNN-LSTM Model. *J. North China Univ. Water Resour. Electr. Power (Soc. Sci. Ed.)* **2023**, *39*, 15–22.
31. Qiu, R.; Zhou, H.; Wu, H. Short-Term Prediction Study of Berth Demand Based on LSTM Recurrent Neural Network. *Technol. Appl. Autom.* **2019**, *38*, 107–113.
32. Chen, Y.; Zhao, H.; Zhou, R.; Xu, P.; Zhang, K.; Dai, Y.; Zhang, H.; Zhang, J.; Gao, T. CNN-BiLSTM Short-Term Wind Power Forecasting Method Based on Feature Selection. *IEEE J. Radio Freq. Identif.* **2022**, *6*, 922–927. [CrossRef]
33. Staffini, A. A CNN–BiLSTM Architecture for Macroeconomic Time Series Forecasting. *Eng. Proc.* **2023**, *39*, 33.
34. Cai, H.; Chen, X.; Ling, J.; Xu, Q. Short-Term Load Forecasting Based on Radam Optimized CNN-BiLSTM-AE Hybrid Model. In Proceedings of the 2022 Power System and Green Energy Conference (PSGEC), Shanghai, China, 25–27 August 2022; pp. 626–631.

35. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
36. Zhang, Y.; Cao, H.; Kan, X. Snow Cover Identification in Xinjiang Region Based on Spatiotemporal Feature Fusion of FY-4A/AGRI. *Remote Sens. Technol. Appl.* **2020**, *35*, 1337–1347.
37. Zhuo, D.; Jing, J.; Zhang, H. Classification of Short Cut Felt Defects Based on Convolutional Neural Networks. *Prog. Lasers Optoelectron.* **2019**, *56*, 144–151.
38. Rumelhart, D.E.; Hintont, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
39. Li, W.; Yang, X.; Chen, K. Automatic Heart Sound Classification Based on CNN and RNN Joint Network. *Comput. Eng. Des.* **2020**, *41*, 46–51.
40. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
41. Jiang, Q.; Feng, R.; Zhang, R.; Wang, J. Multi-Scenario Gait Authentication on Smartphones Based on GRU. *J. Netw. Inf. Secur.* **2022**, *8*, 26–39.
42. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762.
44. Zhou, F.; Zhao, J.; Zhou, C. Determination of the Optimal Sound Speed Formula for Multibeam Echosounder Systems. *Taiwan Strait* **2001**, *4*, 411–419.
45. Li, B.; Zhai, J. A Novel Sound Speed Profile Prediction Method Based on the Convolutional Long-Short Term Memory Network. *J. Mar. Sci. Eng.* **2022**, *10*, 572. [CrossRef]