AIMS Mathematics, 11(7): 22770–22806.
DOI: 10.3934/math.2026919
Received: 25 April 2026
Revised: 30 June 2026
Accepted: 06 July 2026
Published: 28 July 2026

https://www.aimspress.com/journal/Math

Research article

Deep learning for global sea surface temperature anomaly forecasting with balanced spatiotemporal features

Yan Li[1*], **Weiyi Chen**[1,2], **Jia Shi**[1], **and Zixuan Lin**[1]

[1] School of Mathematics and Statistics, Changchun University, Changchun 130012, China
[2] Anyang Wenfeng District Tax Bureau, State Taxation Administration, Anyang 455000, China

* **Correspondence:** Email: yli202601@163.com; Tel: +86-18943090670.

Abstract: With climate change intensifying, sea surface temperature anomalies (SSTAs) have become a key climate indicator. However, their nonlinear, non-stationary, and periodic characteristics pose challenges to traditional forecasting methods. This paper proposes a hybrid forecasting framework that integrates three-level temporal decomposition with deep learning, extending single-point prediction to multi-region spatiotemporal forecasting. First, a hybrid SCV module integrating singular spectrum analysis (SSA), complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), and variational mode decomposition (VMD) is established to disentangle the original signal into trend, periodic, and stochastic components, which are subsequently reconstructed via fuzzy entropy.These components are then fed into deep learning models (e.g., bidirectional gated recurrent unit (BiGRU)) for prediction. Second, we combine temporal features with spatial features to propose the SCV-spatiotemporal feature coupling block (STFCB) model. This model uses a feature pyramid network (FPN) and an intrinsic mode function (IMF)-guided convolutional block attention module (IMF-guided CBAM) to dynamically fuse spatial information. The results demonstrate that SCV-STFCB significantly outperforms the temporal-only SCV with feature attention (SCV-FA) egret swarm optimization algorithm (ESOA) model, with BiGRU achieving a mean square errora mean square error of 0.000073 and R^2 of 0.995903, indicating that the proposed spatiotemporal enhancement method improves the test-period predictive performance.

Keywords: sea surface temperature anomaly prediction; time series decomposition; feature pyramid network; convolutional block attention module; deep learning
Mathematics Subject Classification: 68T07, 62M10,86A10

1. Introduction

The ocean is a central component of the global climate system because it stores and transports heat, regulates air–sea exchange, and affects atmospheric circulation and water vapor supply [1]. Under anthropogenic warming, the upper ocean has absorbed most of the excess heat accumulated in the climate system, making ocean temperature variation a direct signal of climate change [2]. Sea surface temperature (SST) is therefore widely used to monitor oceanic thermal conditions. Compared with absolute SST, sea surface temperature anomalies (SSTAs) remove the climatological mean and are more suitable for identifying abnormal warming, cooling, and interannual variability [3]. Accurate SSTA forecasting is also important for marine ecosystems, fisheries, precipitation anomalies, typhoon activity, and air–sea feedback analysis [4, 5].

Long-term insitu observations and satellite-based remote sensing have provided increasingly complete SST records. Reconstructed datasets such as HadSST and other global SST products support climate analysis, model validation, and long-period forecasting studies [6, 7]. However, SSTA prediction remains difficult. The observed sequence is nonlinear, nonstationary, and multiscale, while gridded SSTA fields contain spatial dependence and regional propagation patterns. Traditional statistical methods are interpretable but limited in representing nonlinear and long-range dependence [8]. Deep learning models improve feature learning, but their performance still depends strongly on whether temporal components and spatial structures are adequately extracted [9].

Existing SST and SSTA forecasting methods can be roughly divided into physical models, statistical models, and data-driven learning models. Physical models describe ocean–atmosphere processes through dynamical and thermodynamic equations. Empirical orthogonal function reconstruction and related methods have improved historical SST analysis and uncertainty estimation [10, 11], while extended reconstructed SST products and optimal interpolation have enhanced spatial consistency in global SST fields [12, 13]. Coupled ocean–atmosphere models have also been used for El Niño–Southern Oscillation (ENSO) and tropical Atlantic-focused prediction because of their physical interpretability [14, 15]. Nevertheless, these models are sensitive to the initial and boundary conditions, and coupled general circulation models usually require a high computational cost for seasonal prediction [16].

Statistical models use historical SST records and climate indicators to build forecasting relationships. Canonical correlation analysis and lagged regression have been applied for global and regional SST prediction [17, 18]. Nawi et al. developed a hybrid autoregressive integrated moving average (ARIMA)-SVM model for SST prediction and confirmed its superiority over standalone ARIMA and SVM models across multiple accuracy metrics [19]. To better handle nonlinear and nonstationary signals, temporal decomposition has been introduced into SST forecasting. Ensemble empirical mode decomposition with autoregressive integrated moving average (EEMD-ARIMA) models extract multiscale components before prediction, and improved empirical mode decomposition combined with long short-term memory (EMD-LSTM) models combine decomposition with neural networks to reduce errors [20, 21]. Improved variational mode decomposition (VMD) and EEMD-XGBoost frameworks also show that decomposition can reduce redundancy and expose useful temporal structures [22, 23]. However, many studies still rely on single or dual decomposition methods. Such designs may not fully separate long-term trends, periodic oscillations, and high frequency stochastic fluctuations.

Deep learning has become an important direction for SST and SSTA prediction because it can learn nonlinear patterns from data. Early neural network studies showed advantages in tropical Pacific SSTA

forecasting [24, 25]. Long short-term memory (LSTM) networks, gated recurrent units (GRUs), and bidirectional recurrent networks are widely used to model temporal dependence [26–28]. Attention mechanisms and convolutional structures further help identify key temporal intervals and spatial patterns [29, 30]. In SST applications, principal component analysis coupled with long short-term memory (PCA-LSTM), long short-term memory enhanced by adaptive boosting (LSTM-AdaBoost), self-organizing map fused with long short-term memory (SOM-LSTM), and related models have improved nonlinear or region-specific prediction [31–35].

Recent research has shifted from purely temporal models to spatiotemporal deep learning. Convolutional neural network (CNN)-LSTM frameworks combining 2D and 3D convolutions for spatial feature extraction have achieved more accurate short- and mid-term SST prediction [36, 37]. Graph convolutional networks describe spatial connections among ocean grid points and are suitable for modeling regional dependence [38, 39]. attention-based models, including hybrid dilated convolution and bidirectional gated recurrent unit with attention mechanism (HDC-BiGRU-AT), attention mechanism with bidirectional long short-term memory (Attention-BiLSTM), and the hierarchical graph recurrent network (HiGRN) framework with adaptive embedding and multilevel attention, have been widely applied to capture both local and global SST patterns and improve feature weighting [40–42]. Graph recurrent attention models and spatiotemporal transformer frameworks further strengthen long-range spatiotemporal interaction modeling [43–45]. Despite these advances, many models rely solely on raw sequential inputs or focus predominantly on spatial features, leaving the interaction between decomposed temporal components and spatial feature selection largely underexplored.

On this basis, the SCV-spatiotemporal feature coupling block (STFCB) model is constructed to incorporate spatial information. The STFCB combines a feature pyramid network (FPN) with an intrinsic mode function (IMF)-guided convolutional block attention module (CBAM). FPN extracts multiscale spatial representations from gridded SSTA fields, while IMF-guided CBAM uses temporal component information to adjust channel and spatial attention weights. In this way, temporal features can inform spatial feature selection, allowing the model to jointly exploit decomposed temporal signals and spatial anomaly structures. To improve the reliability of evaluation, the proposed framework is compared with climate baselines, neural network baselines, and state-of-the-art SSTA prediction models. Multistep prediction, rolling window cross-validation, ablation experiments, and computational cost analyses are also included.

The main contributions of this study are threefold. First, a progressive temporal decomposition framework is developed to separate trend, periodic, and high-frequency components in SSTA forecasting. Second, a spatiotemporal feature fusion framework is proposed by integrating FPN and IMF-guided CBAM, which enhances multiscale spatial representation and attention-based feature selection. Third, a comprehensive experimental design is provided, including traditional climate baselines, mainstream neural networks, state-of-the-art spatiotemporal models, multistep forecasting settings, rolling window validation, and module-level ablation variants.

The remainder of this paper is organized as follows. Section 2 introduces the dataset, terminology, decomposition methods, IMF-guided attention mechanism, evaluation metrics, and experimental settings. Section 3 presents the decomposition results, feature visualization, and prediction results. Section 4 discusses the model comparison, multistep forecasting, stability validation, ablation analysis, and computational cost. Section 5 concludes the study and outlines limitations and future work.

2. Materials and methods

2.1. Nomenclature

To ensure consistency and clarity, the main abbreviations and terminology adopted in this paper are summarized in Table 1. The proposed model consists of multiple modules including decomposition, optimization, the attention mechanism, and spatiotemporal feature fusion. Therefore, clarifying relevant terms before elaborating on the research methodology can prevent ambiguity in the subsequent content.

Table 1. Abbreviations and terminology used in this study.

Abbreviation	Full name	Meaning in this study
SST	Sea surface temperature	Sea surface temperature
SSTA	Sea surface temperature anomaly	Deviation of SST from the climatological mean
SSA	Singular spectrum analysis	First-level temporal decomposition for trend extraction
CEEMDAN	Complete ensemble empirical mode decomposition with adaptive noise	Second-level temporal decomposition for nonlinear residual components
IMF	Intrinsic mode function	Component generated by CEEMDAN
FE	Fuzzy entropy	Complexity measure for IMF reconstruction
FA-IMF	Fuzzy entropy aggregated IMF	Reconstructed IMF group based on fuzzy entropy clustering
VMD	Variational mode decomposition	Third-level decomposition for high-complexity component refinement
ESOA	Egret swarm optimization algorithm	Optimization algorithm for VMD parameter selection
FPN	Feature pyramid network	Multi-scale spatial feature extraction module
CBAM	Convolutional block attention module	Channel and spatial attention module
IMF-guided CBAM	IMF-guided convolutional block attention module	Attention module in which decomposed temporal components guide the channel and spatial attention weights
STFCB	Spatiotemporal feature coupling block	Spatiotemporal feature coupling module that integrates FPN, IMF-guided CBAM, and projection-based feature fusion
SCV-STFCB	SSA-CEEMDAN-VMD with spatiotemporal feature coupling block	Proposed spatiotemporal prediction framework

2.2. Methods

2.2.1. Complete ensemble empirical mode decomposition with adaptive noise

In signal processing, EEMD and CEEMD partially alleviate the mode mixing problem of EMD, but white noise may still remain in the decomposition results, and the introduction of noise can lead to inconsistent numbers of modes. To address these issues, CEEMDAN uses adaptive noise and a complete ensemble strategy, further enhancing its signal decomposition and noise reduction capabilities. Its computational steps are as follows.

(1) Initialize the residual $r_0(t) = x(t)$, where $x(t)$ is the original signal.

(2) Add white noise to the signal to generate and apply EMD to obtain the first intrinsic mode function. Then compute the ensemble average and update the residual

$$\mathrm{IMF}_1(t) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{EMD}_1\big(x(t) + \epsilon_0\, w^{(i)}(t)\big), \tag{2.1}$$

$$r_1(t) = x(t) - \mathrm{IMF}_1(t). \tag{2.2}$$

(3) For $k \geq 1$, add adaptive noise to the residual and apply EMD to obtain the $(k+1)$-th IMF. Then compute the ensemble mean and update the residual

$$\mathrm{IMF}_{k+1}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{EMD}_1\Big(r_k(t) + \epsilon_k\, \mathrm{EMD}_k\big(w^{(i)}(t)\big)\Big), \tag{2.3}$$

$$r_{k+1}(t) = r_k(t) - \mathrm{IMF}_{k+1}(t). \tag{2.4}$$

(4) Repeat Step (3) until the residual $r_k(t)$ becomes a monotonic function or satisfies a predefined stopping criterion. Finally, the original signal can be expressed as the sum of all IMFs and the final residual.

2.2.2. Variational mode decomposition

VMD is an adaptive and nonrecursive signal decomposition method for time-frequency analysis. It formulates the decomposition process as a constrained variational problem, adaptively determines the number of modes according to the signal's characteristics, and iteratively estimates the optimal center frequency and bandwidth for each mode. Its main procedure can be summarized as follows.

(1) Initialize the modal components $\{u_k^1\}$, their center frequencies $\{\omega_k^1\}$, the Lagrange multiplier λ^1, and set the iteration index $n = 0$.

(2) Set $n = n + 1$ and enter the iterative optimization process.

(3) For each $k = 1, 2, \ldots, K$, update the modal component u_k^n and its center frequency ω_k^n using the corresponding update equations.

(4) Update the Lagrange multiplier λ^n using its update rule.

(5) Given a convergence threshold $\varepsilon > 0$, terminate the iteration if the following condition is satisfied:

$$\sum_{k=1}^{K} \frac{\|u_k^n - u_k^{n-1}\|_2^2}{\|u_k^{n-1}\|_2^2} < \varepsilon.$$

Otherwise, return to Step (2) and continue the iterative process.

Here, u_k denotes the decomposed single-component amplitude-modulated frequency-modulated (AM-FM) mode, ω_k represents the center frequency of each mode, λ is the Lagrange multiplier, and n denotes the iteration index.

2.2.3. Bidirectional long short-term memory network

The core idea of bidirectional long short-term memory (BiLSTM) is to model sequential data by leveraging both past and future information simultaneously, consisting of a forward and a backward LSTM. The structure of LSTM is shown in Figure 1. It addresses long-term dependency problems through a gate mechanism that controls the flow of information.

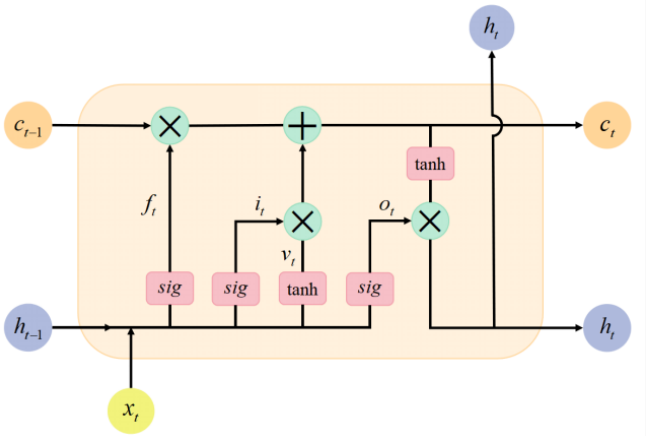


Figure 1. LSTM structure.

The LSTM's architecture includes three gates: The forget gate, the input gate, and the output gate, each using a sigmoid function to regulate information transmission. The output equations of an LSTM unit are as follows:

$$\begin{aligned} i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i),\ f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o),\ \tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t,\ h_t = o_t \odot \tanh(c_t), \end{aligned} \tag{2.5}$$

where x_t is the input at time t, h_t is the hidden state, c_t is the cell state, $\tilde{c}_t$ is the candidate cell state, $\sigma(\cdot)$ is the sigmoid function, $\odot$ denotes element-wise multiplication, and W and b are weight matrices and bias vectors, respectively.

In the BiLSTM model, the forward LSTM processes the input sequence while the backward LSTM processes the reversed sequence. Their outputs are linearly combined to produce the final result. Thus, the forward LSTM predicts future information from historical data, whereas the backward LSTM fits past data using future information. The models architecture is illustrated in Figure 2.

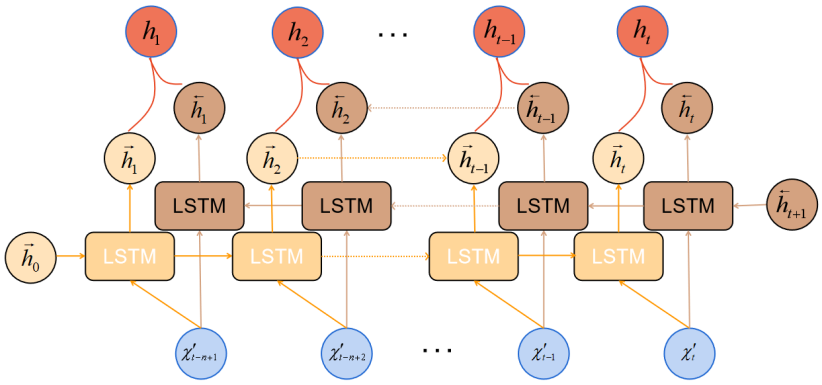


Figure 2. BiLSTM structure diagram.

2.2.4. Bidirectional gated recurrent unit

The bidirectional gated recurrent unit (BiGRU) extends the conventional GRU by introducing a bidirectional structure. It adds a reverse propagation path, which enables the model to extract temporal information from both the forward and backward directions. The GRU enhances computational efficiency by simplifying LSTM's gating mechanism, retaining only the update gate and the reset gate, which significantly reduces the parameter count while still effectively capturing long-term dependencies. Figure 3 illustrates the GRU's architecture.

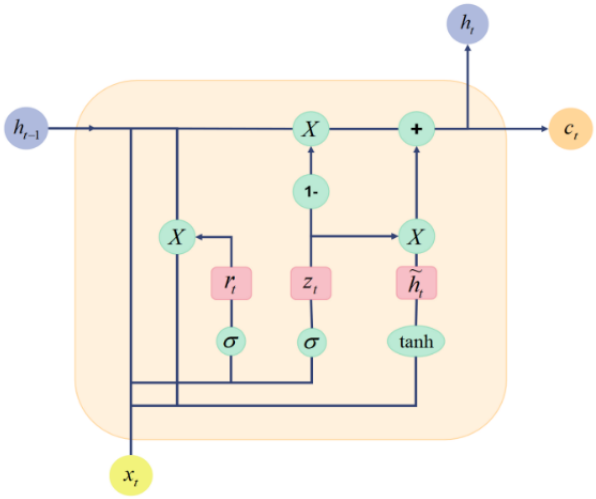


Figure 3. GRU structure diagram.

The GRU's core computational steps are as follows.

(1) The update gate determines the impact of the current input on the hidden state and whether to retain historical information.

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z), \tag{2.6}$$

where z_t is the update gate signal that controls memory retention.

(2) The reset gate determines how much historical information to retain, and its formula is

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r), \tag{2.7}$$

where r_t is the reset signal (larger values indicate more retained information) and W_r is the weight matrix.

(3) The candidate hidden state combines the current input with the reset-gate-adjusted historical information

$$\widetilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t]), \tag{2.8}$$

where $\widetilde{h}_t$ is the candidate hidden state, $\odot$ denotes element-wise multiplication, and tanh normalizes the output.

(4) The final hidden state combines the update gate and the candidate hidden state

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t, \tag{2.9}$$

where h_t is the current hidden state, $(1 - z_t) \odot h_{t-1}$ retains past information, and $z_t \odot \widetilde{h}_t$ incorporates new information, achieving a balance between forgetting and updating.

BiGRU consists of two independent GRU networks processing the input sequence in the forward and backward directions, respectively. The forward GRU generates hidden states $\overrightarrow{h}_t$ and the backward GRU generates $\overleftarrow{h}_t$. The final output is produced by concatenating both hidden states. The architecture of BiGRU is shown in Figure 4.

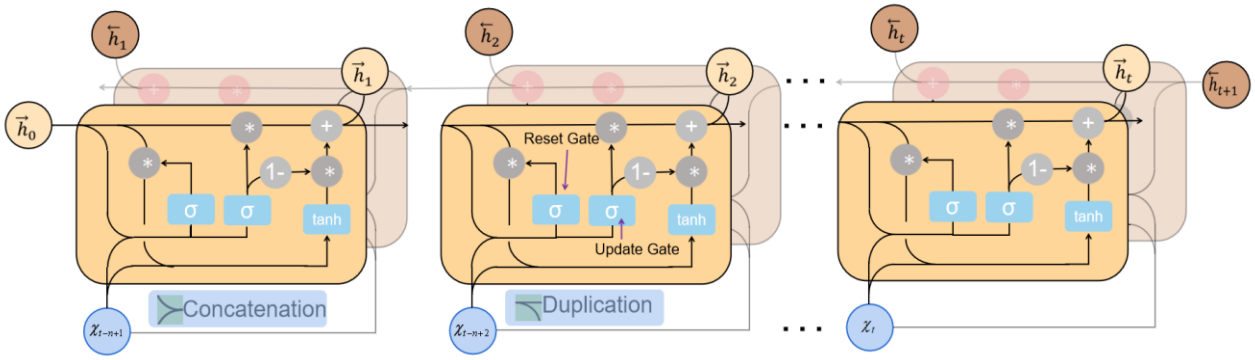


Figure 4. BiGRU structure diagram.

2.2.5. Fuzzy entropy for complexity quantification

FE is a measure of time series complexity that quantifies the degree of irregularity or randomness. Unlike approximate entropy and sample entropy, FE uses a fuzzy membership function to achieve a continuous and smooth similarity measure, making it more robust to noise. In this study, FE is used to evaluate the dynamic complexity of each IMF obtained from the CEEMDAN decomposition, thereby guiding the aggregation of IMFs with similar complexity levels.

Definition of FE: Given a time series of length N, $\{u(i) : 1 \le i \le N\}$, the FE is computed as follows.

(1) Reconstruct the phase space with the embedding dimension m to form vectors

$$X_i^m = \{u(i), u(i+1), \ldots, u(i+m-1)\} - u_0(i), \quad i = 1, 2, \ldots, N-m+1. \tag{2.10}$$

where

$$u_0(i) = \frac{1}{m} \sum_{j=0}^{m-1} u(i+j), \tag{2.11}$$

and $u_0(i)$ is the mean of the i-th reconstructed vector.

(2) Define the distance between two vectors X_i^m and X_j^m as the maximum absolute difference between their corresponding components

$$d_{ij}^m = \max_{k\in[0,m-1]} \left|(u(i+k) - u_0(i)) - (u(j+k) - u_0(j))\right|. \tag{2.12}$$

(3) Compute the similarity degree between the two vectors using an exponential fuzzy membership function

$$D_{ij}^m = \exp\left(-\frac{(d_{ij}^m)^n}{r}\right), \tag{2.13}$$

where r is the similarity tolerance, typically set according to the standard deviation of the time series, and n controls the gradient of the fuzzy boundary. In this study, $n = 2$.

(4) Define the average similarity degree as

$$\phi^m = \frac{1}{N-m}\sum_{i=1}^{N-m}\left[\frac{1}{N-m-1}\sum_{\substack{j=1\\ j\neq i}}^{N-m} D_{ij}^m\right]. \tag{2.14}$$

(5) Repeat Steps (1)–(4) for the embedding dimension $m + 1$ to obtain ϕ^{m+1}.

(6) Finally, FE is defined as

$$\mathrm{FE}(m, r, N) = \ln\phi^m - \ln\phi^{m+1}. \tag{2.15}$$

In this study, the parameters are set as

$$m = 2, r = 0.15 \times \mathrm{std}(u) \textit{ and } n = 2,$$

following standard practice in time series analysis.

Dynamic complexity interpretation: A higher FE value indicates greater irregularity, complexity, and randomness in the signal, typically corresponding to high-frequency noise or stochastic fluctuations. A lower FE value suggests more regular, ordered, and predictable dynamics, often associated with periodic or low-frequency oscillatory components.

FE similarity criteria for IMF aggregation: After computing the FE value for each IMF, we apply K-means clustering to group IMFs with similar FE values. The similarity criterion is based on the Euclidean distance between FE values. The number of clusters K_2 is determined by the elbow method based on the within-cluster sum of squares (WCSS). IMFs within the same cluster are summed to form a fuzzy entropy aggregation component (FA-IMF). This aggregation strategy reduces the number of input components while preserving the distinct dynamic characteristics of low-complexity periodic signals and high-complexity stochastic signals.

Pseudocode for FE-based IMF aggregation.

Algorithm 1 FE-based IMF aggregation

Require: IMF list $\{IMF_1, IMF_2, \ldots, IMF_{K_1}\}$ obtained from CEEMDAN
Ensure: FA-IMF list $\{FA\text{-}IMF_1, FA\text{-}IMF_2, \ldots, FA\text{-}IMF_{K_2}\}$
for each IMF_i **do**
  Compute FE_i using $m = 2$ and $r = 0.15 \times \mathrm{std}(IMF_i)$
end for
Normalize the FE values

$$\widetilde{\mathrm{FE}}_i = \frac{\mathrm{FE}_i - \min(\mathrm{FE})}{\max(\mathrm{FE}) - \min(\mathrm{FE})}$$

Determine the optimal number of clusters K_2 using the elbow method
Apply K-means clustering to $\{\widetilde{\mathrm{FE}}_i\}$ to obtain cluster assignments $C_1, \ldots, C_{K_2}$
for $j = 1$ to K_2 **do**
  $FA\text{-}IMF_j = \sum_{i \in C_j} IMF_i$
end for
return $FA\text{-}IMF_1, \ldots, FA\text{-}IMF_{K_2}$

2.3. SCV-STFCB model

2.3.1. Data and processing

The dataset used in this study is HadSST.4.2.0.0, a SST observation dataset released by the Met Office Hadley Centre. This dataset is available for download under the Open Government Licence v3 from the Met Office website at https://www.metoffice.gov.uk/hadobs/hadsst4/. The HadSST.4.2.0.0 dataset contains a collection of 200 elements, all of which have been corrected for systematic SST errors. The anomalies are calculated relative to the 1961–1990 average and correspond to SSTAs at a depth of approximately 20 cm.

The HadSST.4.2.0.0 dataset provides monthly, globally averaged SSTAs from 1850 to the present, obtained by weighted averaging of global observation grid points. The SSTAs are computed relative to the 1961–1990 climatological baseline and expressed in degrees kelvin (K), equivalent to degrees Celsius (°C). For the proposed SCV module, consecutive monthly observations from January 1850 to December 2024 are used. To prevent information leakage, the time series was chronologically divided into training and testing sets at an 8:2 ratio. The min-max normalization parameters were computed exclusively from the training set, and the resulting normalization was subsequently applied to both the training and testing sets, as formulated below:

$$X_{\mathrm{scaled},i} = \frac{X_i - X_{\min}^{\mathrm{train}}}{X_{\max}^{\mathrm{train}} - X_{\min}^{\mathrm{train}}} \tag{2.16}$$

where X_i denotes an original value from either the training set or the testing set, and $X_{\min}^{\mathrm{train}}$ and $X_{\max}^{\mathrm{train}}$ denote the minimum and maximum values of the training set, respectively.

For the proposed STFCB module, which targets the spatial correlation of SSTA, the original spatial grid data from the HadSST4.2.0.0 dataset are used. The data are provided on an equal-latitude–longitude grid with a resolution of $5^\circ \times 5^\circ$, comprising 36 latitude grid points (90°S to 90°N) and 72 longitude grid points (0° to 358°E). Each month's observation corresponds to a two-dimensional global

grid field. When stacked along the temporal dimension, these observations form a three-dimensional spatiotemporal structure with the dimensions (time, latitude, longitude). The anomaly value for each grid cell is calculated from all available observations within that cell and month; cells without observations are marked as "no data". This three-dimensional structure serves as the direct input to the STFCB module.

The results from the statistical analysis of the grid dataset are presented in Table 2. In terms of sample size, the dataset contains approximately 2.02 million valid data points, indicating extensive temporal coverage and broad spatial extent, which adequately reflect the spatiotemporal variation characteristics of SST in the study region. Regarding central tendency, the mean is 0.0288 and the median is 0.0211. The close proximity of these two values suggests a relatively symmetric overall data distribution. In terms of dispersion, the standard deviation is approximately 0.9809, indicating that SST exhibits some fluctuations across different temporal and spatial scales, yet the overall variation remains within a reasonable range. These statistics demonstrate that the data are in a relatively stable state over the long term, making them suitable as a foundation for subsequent modeling and analysis.

Table 2. Statistical summary of SSTA grid data.

Statistical indicator	Value
Count	2020438
Mean	0.028806
Std	0.980945
Min	-8.246538
25%	-0.512918
50%	0.021054
75%	0.561539
Max	8.556515

To impute missing values in the SSTA grid dataset, a convolutional autoencoder (CAE) with residual connections is constructed. The CAE is an unsupervised neural network that integrates CNN-based feature extraction with data reconstruction capability, making it especially suitable for image-like grid data. Its computational procedure is as follows.

Step 1: Preprocess the raw data, which contain three dimensions: Time, latitude, and longitude. Define a mask matrix

$$M_{t,i,j} = \begin{cases} 1, & X_{t,i,j} \neq \text{NaN} \\ 0, & X_{t,i,j} = \text{NaN} \end{cases} \tag{2.17}$$

Step 2: Extract spatial features using an encoder, then reconstruct the data with a decoder. Residual connections are added to assist learning and improve the reconstruction accuracy. A mask-weighted MAE loss is designed, where errors are computed only over valid regions. The masked MAE is defined as

$$L_{MAE} = \frac{\sum_{t,i,j} M_{t,i,j} \cdot \left| Y_{t,i,j} - X^{norm}_{t,i,j} \right|}{\sum_{t,i,j} M_{t,i,j}}. \tag{2.18}$$

Step 3: To avoid oversmoothing, a gradient constraint term is introduced to preserve spatial structural information

$$L_{grad} = \mathbb{E}\left[|\nabla_x Y| + \left|\nabla_y Y\right|\right], \tag{2.19}$$

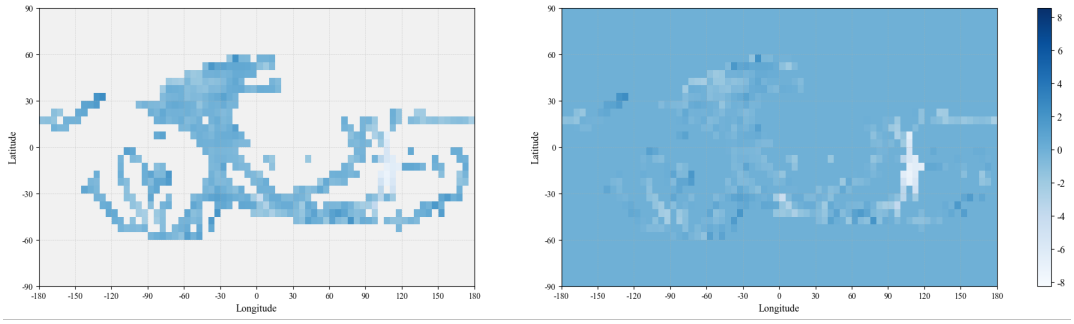
where $\nabla_x Y = Y_{i,j} - Y_{i,j+1}$ and $\nabla_y Y = Y_{i,j} - Y_{i+1,j}$.

Step 4: The total loss function is defined as

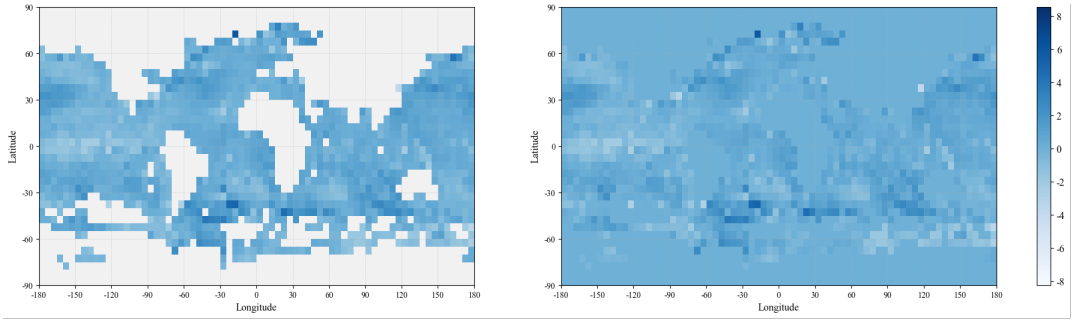
$$L = L_{MAE} + \lambda L_{grad}, \tag{2.20}$$

where λ is the weight coefficient, set to 0.1 in this study. After model training, only the originally missing positions are filled with the predicted values, yielding the complete final dataset.

Statistical results show that after imputation, the mean decreases from 0.0288 to 0.0106, and the standard deviation decreases from 0.9818 to 0.5978, indicating reduced dispersion and a smoother overall distribution. Figure 5 visually compares the imputation effects, with Panels (a) and (b) corresponding to spatial distributions at t=100 and t=1800, respectively. In the raw data, large land and missing regions appear in gray, resulting in a highly discontinuous spatial distribution. After CAE imputation, the missing regions are fully reconstructed, producing a continuous and smooth spatial field, with the oceanic structural features restored. Furthermore, comparing the two time points, the raw valid data at t=1800 are noticeably more abundant than those at t=100, reflecting the continuous advancement of ocean observation technology and the gradual improvement in global ocean coverage over time.



(a)



(b)

Figure 5. Comparison of CAE-based imputation.

To further validate the reliability of the CAE-based imputation, a masked observation experiment was conducted. Specifically, 5%, 10%, and 20% of known valid SST observations in the training period's gridded data were randomly masked. The trained CAE model was then used to reconstruct these masked values, and the reconstructed values were compared with the original observations using root mean square error (RMSE), mean absolute error (MAE), and bias. This validation experiment was designed to examine whether the CAE model can recover missing values without introducing systematic deviation.

As shown in Table 3, the CAE reconstruction performance remains stable under different masking ratios. The RMSE is approximately 0.978 K, and the MAE is approximately 0.711 K across the three masking settings. In addition, the bias remains close to zero, ranging from -0.0039 K to -0.0003 K, indicating that the CAE-based imputation does not introduce obvious systematic deviation. Although the RMSE is comparable with the standard deviation of the original SSTA, this is reasonable because the CAE mainly preserves large-scale spatial structures, while individual grid-point values are affected by natural SST variability. Overall, the masked observation experiment suggests that the CAE-based imputation can provide relatively stable reconstruction performance under the tested masking ratios and does not show obvious systematic bias. However, the reconstruction mainly preserves large-scale spatial structures, while local grid-point errors may still exist due to natural SST variability and sparse observations.

Table 3. CAE imputation performance metrics.

Mask	Ratio	RMSE (K)	MAE (K)	Bias (K)
0	5%	0.9782	0.7119	-0.0039
1	10%	0.9776	0.7111	-0.0009
2	20%	0.9797	0.7112	-0.0003

2.3.2. SCV-STFCB model construction

Given the temporal features extracted in the previous sections, this section introduces the SCV-STFCB model. The model combines spatial feature extraction with spatiotemporal feature fusion. It is designed to capture both temporal evolution and spatial dependence in SSTA fields.The specific model structure is shown in Figure 6. This model integrates multiple decomposition techniques with a spatiotemporal feature extraction module to jointly incorporate spatial information, thereby enhancing the prediction accuracy of the neural network. The model comprises two core components: The temporal decomposition module SCV, where S denotes singular spectrum analysis (SSA), C denotes CEEMDAN, and V denotes VMD; and the spatiotemporal feature extraction module STFCB, which integrates the FPN with the IMF-guided CBAM. The SCV module takes time series SSTA data as input, while the STFCB module models gridded spatial data. Finally, a deep neural network outputs the temperature anomaly predictions. By fully exploiting the multilevel spatiotemporal information inherent in SSTA data, the model enhances its spatiotemporal awareness for prediction tasks.

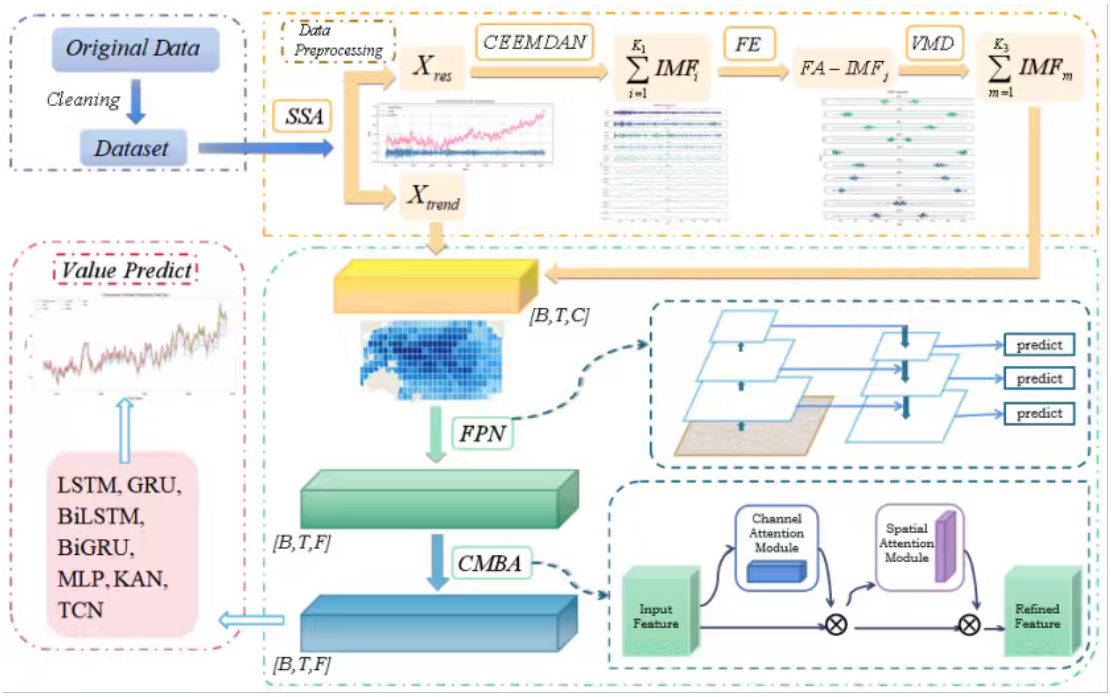


Figure 6. Structural diagram of the SCV-STFCB model.

Let the study region be defined as a spatial grid of size $G = \{(x_i, y_i) | i = 1, \ldots, I, j = 1, \ldots, J\}$, and let the time window be of length $t = 1, ..., T$. The SSTA field is defined as a tensor $X \in \mathbb{R}^{T \times I \times J}$, where $X_{t,i,j}$ denotes the SSTA value at time t and the grid point (i, j). The core mathematical formulation of the SCV-STFCB model can be summarized as

$$\widehat{X}_{t+H} = P\left(\mathbf{Z}_{\text{fuse}}\right), \qquad \mathbf{Z}_{\text{fuse}} = \phi_f\left([\mathbf{Z}_T; \mathbf{Z}_S]\right) = \delta\left(\mathbf{W}_f\left[\mathbf{Z}_T; \mathbf{Z}_S\right] + \mathbf{b}_f\right). \tag{2.21}$$

$$\mathbf{Z}_T = \varepsilon_T\left(X_{t-L+1:t}\right), \qquad \mathbf{Z}_S = \text{GAP}\left(F_{\text{out}}\right), \qquad F_{\text{out}} = \varepsilon_S\left(X_{t-L+1:t}, G, I(X_{t-L+1:t})\right), \tag{2.22}$$

where $\mathbf{Z}_T$ denotes the temporal feature vector generated by the SCV temporal decomposition branch, and F_{out} denotes the spatially refined feature map produced by the STFCB module. GAP$(\cdot)$ represents global average pooling, which converts F_{out} into the spatial feature vector $\mathbf{Z}_S$. $\phi_f(\cdot)$ is a learnable projection function, $\mathbf{W}_f$ and $\mathbf{b}_f$ are trainable parameters, and $\delta(\cdot)$ denotes a nonlinear activation function. Therefore, the feature fusion operation in SCV-STFCB is explicitly implemented as concatenation followed by learnable projection, rather than an undefined feature addition or simple averaging operation.

To strictly prevent information leakage, all preprocessing procedures including min-max normalization; signal decomposition via SSA, CEEMDAN, and VMD; FE-based grouping; and parameter optimization are performed exclusively on the training set $\mathbf{X}_{\text{train}}$. All decomposition parameters such as the number of IMFs K_1, the clustering centroids C_j, and the number of VMD modes K_3 are estimated solely from the training data. The testing set $\mathbf{X}_{\text{test}}$ is subsequently transformed using the fixed parameters derived from the training set and does not participate in any parameter learning or optimization process. This guarantees that the testing set remains entirely unseen during the entire training phase, thereby ensuring the validity and credibility of the forecasting evaluation.

The specific prediction steps are as follows.

Step 1: The original time series X is decomposed by SSA into a trend component X_{trend} and a residual component X_{res}

$$X = X_{trend} + X_{res}. \tag{2.23}$$

Since X_{trend} represents long-term variations, it is fed directly into a neural network for prediction.

Step 2: The residual component$X_{\text{res}}$is further decomposed using CEEMDAN to obtain a set of IMFs and a residual

$$X_{res} = \sum_{i=1}^{K_1} IMF_i + R_{CEEMDAN}, \tag{2.24}$$

where K_1 is the number of IMFs.

Step 3: The FE of each IMF is computed, and similar IMFs are aggregated into FE aggregated IMFs (FA-IMFs)

$$FA - IMF_j = \sum_{k \in C_j} IMF_k, j = 1, 2, \ldots K_2, \tag{2.25}$$

where C_j denotes the j-th cluster of IMFs with similar FE values. The FA-IMF with the highest FE (the most complex component) is further decomposed using VMD

$$\text{FA-IMF}_{\max} = \sum_{m=1}^{K_3} \text{VMD}_m + R_{\text{VMD}}, \tag{2.26}$$

where K_3 is the number of VMD modes.

Step 4: Each decomposed subseries including trend, FA-IMFs, and VMD modes is fed into a neural network model (LSTM, GRU, BiLSTM, etc.) for individual prediction.

Step 5: To incorporate spatial dependencies, the decomposed components themselves are used as inputs to a FPN to extract multiscale frequency-domain features. These features are then adaptively fused through a CBAM, capturing component-level correlations and dynamically weighting important patterns. The spatially enhanced features are subsequently fed into multiple neural networks for prediction, integrating temporal and component-level information across spatial locations.

Step 6: The model is trained by minimizing the prediction error between the forecasted and actual SSTAs, typically using a loss function such as MSE or mean absolute percentage error (MAPE)

$$\min_{\theta} \frac{1}{H} \sum_{t=T+1}^{T+H} \left(x_t - \widehat{x_t}\right)^2, \tag{2.27}$$

where θ represents the neural network parameters, H is the prediction horizon (stride length), x_t is the true SSTA, and $\widehat{x_t}$ is the predicted value.

Based on the steps above, the basic algorithm of this model is presented as Algorithm 2.

Algorithm 2. SCV-STFCB Model

Require: Training datasets $X_1 = \{x_1, x_2, \ldots, x_T\}$, $X_2 \in \mathbb{R}^{T\times H\times W}$, and maximum iteration count *MaxEpoch*

Ensure: Trained model parameters θ and prediction results $\widehat{X}$

Define the function SCV-STFCB(X)
Initialize model parameters θ
for epoch = 1 to *MaxEpoch* **do**
  for each sample $x \in X_1$ **do**
    Decompose X_1 using SSA to obtain X_{trend} and X_{res}
    Decompose X_{res} using CEEMDAN to obtain $\{IMF_i\}$
    for each IMF_i **do**
      Calculate its fuzzy entropy FE_i
    end for
    Cluster $\{IMF_i\}$ according to fuzzy entropy to obtain $\{FA\text{-}IMF_j\}$
    Select the $FA\text{-}IMF$ component with the maximum fuzzy entropy
    Apply VMD to the selected component to obtain $\{VMD_m\}$
    Construct the temporal component set $T = \{X_{\text{trend}}, FA\text{-}IMF_j, VMD_m\}$
    Extract spatial features F_{fpn} from X_2 using FPN
    Input T and F_{fpn} into CBAM to obtain F_{out}
    Input F_{out} into the prediction network to obtain $\widehat{X}$
    Calculate the loss between $\widehat{X}$ and X_1
    Update model parameters θ through backpropagation
  end for
end for
return θ and $\widehat{X}$

2.3.3. Rationale for the key components of the SCV-STFCB framework

(1) Necessity of three-level decomposition

The three-level decomposition framework is designed according to the heterogeneous structure of SSTA signals rather than as a simple stacking of multiple decomposition algorithms. A monthly SSTA series usually contains a slowly varying climate trend, medium-frequency periodic or quasi-periodic oscillations, and high-frequency irregular fluctuations. Therefore, the original SSTA series can be conceptually expressed as

$$X(t) = T(t) + P(t) + Q(t) + N(t), \tag{2.28}$$

where $T(t)$ denotes the low-frequency trend, $P(t)$ denotes relatively regular periodic variations, $Q(t)$ denotes quasi-periodic or nonlinear oscillations such as ENSO-related variability, and $N(t)$ denotes high-frequency stochastic fluctuations or noise.

If these components are directly fed into a neural network, the model needs to learn the long-term

trend evolution, periodic oscillations, and irregular disturbances simultaneously. This increases the difficulty of model training and may lead to over-smoothed predictions or larger errors around turning points. Therefore, decomposition is introduced to reduce the complexity of the original SSTA series before prediction.

A single decomposition method is not sufficient for this mixed signal. For example, CEEMDAN can decompose nonlinear fluctuations into IMFs but it does not explicitly separate the long-term trend before residual decomposition. As a result, trend information may be distributed across several IMFs, which weakens the interpretability and stability of the decomposed components. Similarly, using VMD alone requires selecting the number of modes and penalty parameters in advance, and its performance may be affected when the input sequence contains both slow trends and high-frequency fluctuations.

A dual decomposition strategy such as CEEMDAN–VMD can further refine high-frequency residual components, but it still lacks an explicit trend extraction stage. In long-term SSTA forecasting, the low-frequency warming trend is an important signal that should be modeled separately. If the trend remains mixed with residual oscillations, VMD may spend part of its modes describing slow variations rather than refining high-complexity components. This reduces the efficiency of decomposition and may weaken subsequent neural network predictions.

For this reason, this study adopts a progressive "trend–residual–complexity–frequency refinement" strategy. First, SSA is used to extract the dominant low-frequency trend component $\hat{T}(t)$, which can be regarded as a low-rank approximation of the original series. The residual component is then obtained as

$$X_{\mathrm{res}}(t) = X(t) - \hat{T}(t). \tag{2.29}$$

Second, CEEMDAN is applied to $X_{\mathrm{res}}(t)$ to obtain IMFs with different time-scale characteristics. This step focuses on nonlinear residual fluctuations after the major trend has been removed. Third, FE is used to quantify the complexity of each IMF. IMFs with similar complexity are aggregated into FA-IMF components, which reduces redundant input while preserving the differences between regular and irregular components. Finally, VMD is applied only to the high-complexity FA-IMF, and the egret swarm optimization algorithm (ESOA) is used to optimize its key parameters. Therefore, VMD is not blindly applied to the entire original series but is used specifically to refine the most irregular component.

The VMD refinement of the high-complexity component can be formulated as the following constrained variational problem:

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}, \tag{2.30}$$

subject to

$$\sum_k u_k(t) = f(t), \tag{2.31}$$

where $u_k(t)$ denotes the k-th decomposed mode, ω_k denotes its center frequency, and $f(t)$ denotes the high-complexity FA-IMF selected by FE. This formulation enables VMD to obtain narrow-band modes by adaptively estimating the bandwidths and center frequencies. In the proposed framework, this refinement is applied only after SSA trend extraction, CEEMDAN residual decomposition, and FE-based complexity identification. Therefore, each stage has a clear functional role: SSA extracts the trend, CEEMDAN decomposes the nonlinear residuals, FE identifies the components' complexity, and

VMD refines high-frequency irregular fluctuations. This design reduces unnecessary decomposition of stable components and focuses additional processing on the components that are most difficult to predict.

To validate the effectiveness of the ESOA in optimizing VMD parameters, we compare it with two widely used optimization algorithms: The genetic algorithm (GA) and particle swarm optimization (PSO). All three algorithms are applied to optimize the number of modes K and the penalty parameter α in the VMD decomposition of the high-complexity FA-IMF component. The prediction backbone is fixed as SCV-STFCB-BiGRU, and the same training/testing protocol is adopted.

Table 4. Performance comparison of different VMD optimization algorithms.

Optimization algorithm	RMSE	MAE	R^2	sMAPE (%)	Convergence speed (iterations)
GA	0.0112	0.0093	0.9945	1.32	45
PSO	0.0108	0.0089	0.9950	1.24	38
ESOA (proposed)	**0.0097**	**0.0079**	**0.9966**	**1.10**	**28**

The results indicate that ESOA achieves the lowest RMSE and sMAPE, the highest R^2, and the fastest convergence among the three algorithms. Since SSTA values frequently fluctuate around zero, the denominator of MAPE can easily approach zero, causing the error metric to become infinite or undefined. In contrast, sMAPE remains stable near zero and imposes symmetric penalties on overestimation and underestimation, making it more reliable and physically reasonable for evaluating anomaly field predictions where both positive and negative departures are equally important. This demonstrates that ESOA is more effective in navigating the VMD parameter space, avoiding local optima, and improving the decomposition quality for nonstationary SSTA signals.

(2) Motivation for IMF-guided CBAM

After clarifying the necessity of the temporal decomposition framework, the following part further explains how the decomposed temporal components guide spatial feature extraction in the STFCB module.

The previous section introduced the overall SCV-STFCB architecture, which integrates temporal decomposition with spatial feature extraction. Central to the STFCB module is the IMF-guided CBAM, which differs from a standard CBAM by leveraging temporal IMFs to modulate both channel and spatial attention. The complete mathematical formulation of this novel attention mechanism is presented below.

In a standard CBAM, channel and spatial attention weights are generated only from the input feature map. Therefore, the attention mainly reflects spatial features' saliency. In contrast, the proposed IMF-guided CBAM introduces decomposed temporal frequency information as an additional guidance signal. The IMF information does not replace the original CBAM structure but modulates its channel and spatial attention weights, allowing the temporal components to guide spatial feature selection.

Let $F_{\text{in}} \in \mathbb{R}^{C\times H\times W}$ be the input feature map from a decomposed component (e.g., an FA-IMF). Let $\mathcal{I} = \{IMF_1, \ldots, IMF_K\}$ be the set of IMFs extracted from the same spatial grid point's temporal signal.

Channel attention:

$$A_c\left(F_{\text{in}}, \mathcal{I}\right) = \sigma\Bigg(W_2 \cdot \delta\left(W_1 \cdot \frac{1}{HW}\sum_{i,j} F_{\text{in}}(i,j)\right) + \sum_{k=1}^{K} \alpha_k \cdot \text{GAP}(IMF_k)\Bigg). \tag{2.32}$$

where α_k are learnable IMF-guided weights, GAP$(\cdot)$ denotes global average pooling, $\delta(\cdot)$ denotes the rectified linear unit (ReLU) activation function, and $\sigma(\cdot)$ denotes the sigmoid activation function.

Spatial attention:

$$A_s\left(F_{\text{in}}, \mathcal{I}\right) = \sigma\left(\text{Conv}_{7\times7}\left[\text{AvgPool}\left(F_{\text{in}}\right); \text{MaxPool}\left(F_{\text{in}}\right); \beta \cdot \text{Std}\left(F_{\text{in}} \odot \mathcal{I}\right)\right]\right) \tag{2.33}$$

where β is a learnable scalar modulating the IMF-guided spatial modulation.

Final output:

$$F_{\text{out}} = A_s\left(F_{\text{in}}, \mathcal{I}\right) \odot \left[A_c\left(F_{\text{in}}, \mathcal{I}\right) \odot F_{\text{in}}\right]. \tag{2.34}$$

This design differs from a standard CBAM and conventional decomposition–deep learning hybrid models in two aspects. First, standard CBAM calculates the channel and spatial attention weights only from the input feature map, whereas the proposed IMF-guided CBAM introduces decomposed temporal components into both attention branches. Second, many hybrid forecasting models use decomposed components only as separate prediction inputs or concatenate them with spatial features at the final stage. In contrast, the proposed SCV-STFCB framework uses IMF information to modulate the attention weights before feature fusion. Therefore, temporal decomposition information not only provides additional input features but also directly affects which spatial channels and grid regions are emphasized by the network.

2.3.4. Evaluation metrics

To evaluate the predictive performance of the models, this study uses five metrics: MSE, MAE, MAPE, and the coefficient of determination (R^2).The formulas for each metric are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{2.35}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \tag{2.36}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{2.37}$$

$$sMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\left(|y_i| + |\hat{y}_i|\right)/2} \tag{2.38}$$

$$\text{Skill score} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{\text{clim}}\right)^2} \tag{2.39}$$

where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, $\bar{y}$ represents the mean of the actual values, and n represents the sample size.

The RMSE is the square root of the average squared differences between the predicted values and the actual values. It reflects the overall magnitude of the prediction errors and assigns a greater penalty to large deviations, making it more sensitive to outliers. MAE is the average of the absolute differences between the predicted values and the actual values, directly reflecting the average magnitude of prediction errors. The smaller the RMSE and MAE values, the higher the prediction accuracy of the model. The core idea of sMAPE is to calculate the absolute difference between each predicted value and its corresponding actual value and normalize it by the average magnitude of the predicted and actual values. A smaller sMAPE indicates a lower relative prediction error and better model performance. R^2 measures how well the model explains the variation in the observed data; the closer R^2 is to 1, the better the model fits the data. The skill score evaluates the improvement of the prediction model relative to the climatology reference model. A skill score close to 1 indicates that the model performs substantially better than the reference model, a value of 0 indicates equivalent performance, and a negative value indicates that the model performs worse than the reference model.

2.3.5. Rolling window cross-validation

To ensure robust evaluation and avoid overfitting to specific temporal periods, we use rolling window cross-validation. To evaluate the model's robustness across different temporal periods, a rolling window cross-validation scheme was adopted. The full dataset spans from 1850 to 2024, covering 174 years. For each fold, a 20-year window was used for training, followed by a 5-year window for validation and a 10-year window for testing. With a stride of 10 years, five folds were generated, each testing on a distinct decadal period (e.g., 1875–1885, 1905–1915, 1935–1945, 1965–1975, 1995–2005). For fold i, the training period is $[t_i, t_i + 20]$ years, the validation period is $[t_i + 20, t_i + 25]$ years, and the testing period is $[t_i + 25, t_i + 35]$ years, with a stride of 10 years. The final reported performance is the average across all five folds. This rolling-window design is adopted because the SSTA record is a long chronological climate sequence rather than an independently and identically distributed dataset. Random cross-validation may mix earlier and later climate regimes and introduce unrealistic temporal information leakage. In contrast, the rolling window strategy preserves the chronological order of the data and evaluates the model on multiple nonoverlapping future periods. Therefore, it provides a more appropriate assessment of whether the model can generalize across different historical climate regimes.This design ensures that models are evaluated on multiple nonoverlapping temporal segments, assessing their ability to generalize across different climate regimes.

We set three prediction horizons: H =3 (3-months ahead), H =6 (6-months ahead), and H =12 (12-months ahead). The input time step is fixed at L =12, meaning that data from the previous 12 months are used to predict the SSTA for the next months (direct multistep strategy). This allows an evaluation of medium- to long-term predictive capability, and for this Algorithm 3 is used.

Algorithm 3 SCV-STFCB (BiGRU) with multistep prediction and rolling window cross-validation

Require: Training dataset $X = \{x_1, x_2, \ldots, x_T\}$, $N_{\text{folds}} = 5$, $H \in \{3, 6, 12\}$, $L = 12$
Ensure: Average evaluation metrics across all folds and horizons
Initialize results[ ]
for fold = 1 to N_{folds} **do**
  Split data into training, validation, and test sets using a rolling window
  for each $h \in H$ **do**
    Train the model with input length L and output horizon h
    Validate the model on the validation set
    Evaluate the model on the global and regional test datasets
    Store MSE, MAE, MAPE, and R^2
  end for
end for
return average metrics across all folds and horizons

2.4. Experimental setup and implementation details

All experiments were conducted on a computing server with an NVIDIA Tesla V100 (32 GB) graphical processing unit (GPU), an Intel Xeon Gold 6248 central processing unit (CPU), and 128 GB of random access memory (RAM). The deep learning models were implemented using PyTorch 2.0.1 with CUDA 11.8.

Hyperparameters: For all neural network models (LSTM, GRU, BiLSTM, BiGRU, multilayer perceptron (MLP), kolmogorov-arnold network (KAN), and temporal convolutional network (TCN)), the input sequence length was set to $L = 12$ (months), and the prediction horizon H was set to 1, 3, 6, or 12, depending on the experiment. The hidden state dimension was set to 64 for all recurrent models (LSTM, GRU, BiLSTM, BiGRU). The MLP consisted of two hidden layers with 128 and 64 neurons, respectively. The KAN model used a layer width of 64 with cubic B-spline basis functions. The TCN used four dilated convolutional layers with a kernel size of 3 and the dilation rates $[1, 2, 4, 8]$. The batch size was set to 32, and the learning rate was initialized at 0.001 with the Adam optimizer. A learning rate scheduler (ReduceLROnPlateau) was applied with a patience of 10 epochs and a reduction factor of 0.5. All models were trained for a maximum of 100 epochs with early stopping triggered if the validation loss did not improve for 20 consecutive epochs. The MSE was used as the primary loss function. For regularization, dropout with a rate of 0.2 was applied to all recurrent and fully connected layers. For the VMD decomposition, the penalty parameter α was set to 2000, the time step τ to 0, and the number of modes K was optimized by the ESOA in the range of $[3, 12]$. For the CBAM module, the reduction ratio for channel attention was set to 16.

All neural-network-based experiments were independently repeated five times using random seeds of 41, 42, 43, 44, and 45. The evaluation metrics are reported as the mean ± standard deviation across the five repeated runs to reduce the influence of random initialization, minibatch sampling, and stochastic optimization. For reproducibility, all compared neural network models were trained under the same data split, input sequence length, prediction horizons, optimizer, learning rate scheduler, maximum epoch number, early stopping criterion, dropout rate, and evaluation metrics.

3. Results

To determine whether the SSTA sequence contains a long-term slow trend, a 36-month moving average and the cumulative sum of deviation metrics were used. The cumulative sum of deviation was calculated by accumulating the de-meaned sequence, with the results shown in Figure 7. The cumulative plot reveals that the data neither exhibit zero mean random fluctuations nor oscillate around a horizontal line. Instead, a distinct pattern of an initial decline followed by recovery emerges, indicating a persistent long-term gradual trend in the SSTA data. Accordingly, the SSA method was adopted for trend separation.

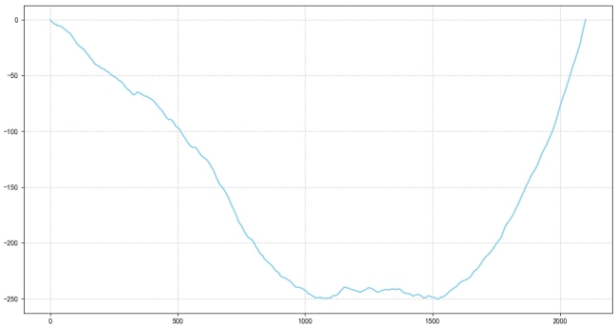


Figure 7. Cumulative chart of the long-term trends of SSTA.

SSA decomposes the original data into a trend component and a residual component, as illustrated in Figure 8. The trend component is smoother than the residual component and exhibits a clear upward trend, which is consistent with the observed fact that the ocean continuously absorbs heat, and SST rises steadily under the background of global warming. In contrast, the residual component fluctuates violently and oscillates frequently around zero, reflecting extreme events and higher-frequency stochastic disturbance information.

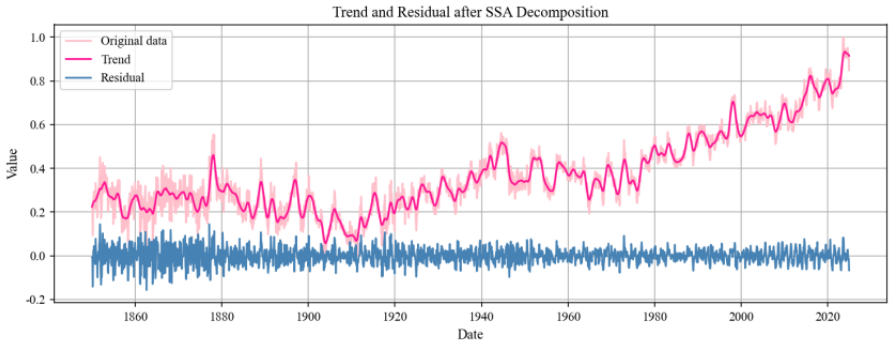


Figure 8. SSA decomposition of SSTA.

Because the high frequency and strong fluctuations of the time series negatively affect the prediction accuracy of machine learning models, this study further extracted information from the residual component after SSA decomposition using CEEMDAN. The decomposition results are shown in Figure 9. CEEMDAN effectively alleviates the mode mixing problem by adding adaptive noise and decomposes the residual component into multiple intrinsic mode functions and one residual term. Each IMF represents the fluctuation characteristics of the time series at different scales or frequencies.

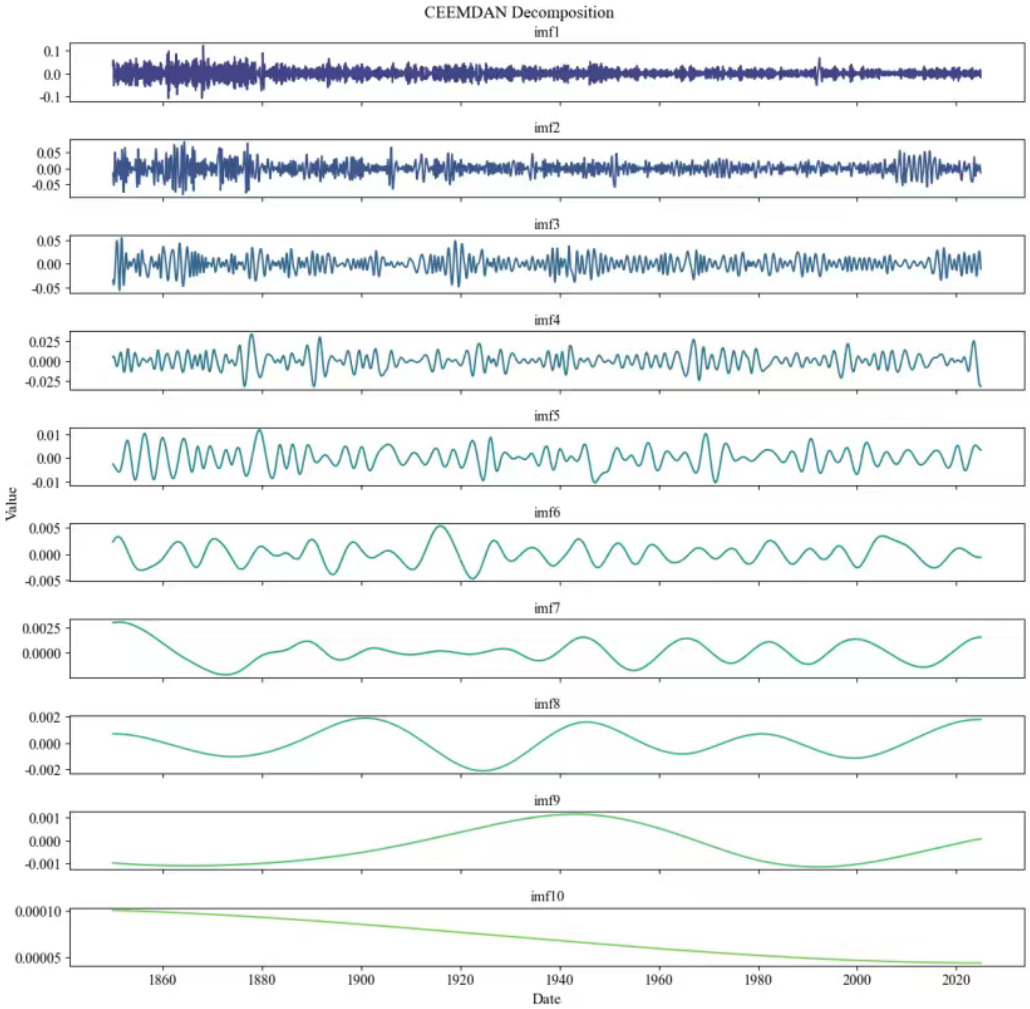


Figure 9. CEEMDAN decomposition of the SSA residual.

The FE value of each IMF was calculated, and IMFs with similar FE values were clustered. The principle of FE calculation is as follows. A high FE value of an IMF indicates that the component has high complexity and may contain more irregular fluctuations or noise. A low FE value indicates that the component is relatively orderly and regular in its variation, which may correspond to periodic components such as seasonal changes. In this study, FE was calculated for eight IMFs, and these IMFs were clustered according to their FE values. IMF1 through IMF4 were grouped into one cluster, and IMF5 through IMF8 were grouped into another cluster. The sum of IMF1 to IMF4 is denoted as FA-IMF1, and the sum of IMF5 to IMF8 is denoted as FA-IMF2. FA-IMF1 has higher FE and contains more complex information. Therefore, FA-IMF1 was further decomposed using VMD, with the decomposition results shown in Figure 10. The parameters α, τ, and K of VMD were optimized using the ESOA and set to 1.06, 0.0036, and 9, respectively. Through these steps, 12 important components were obtained: IMF1 to IMF10, FA-IMF2, and the trend term.

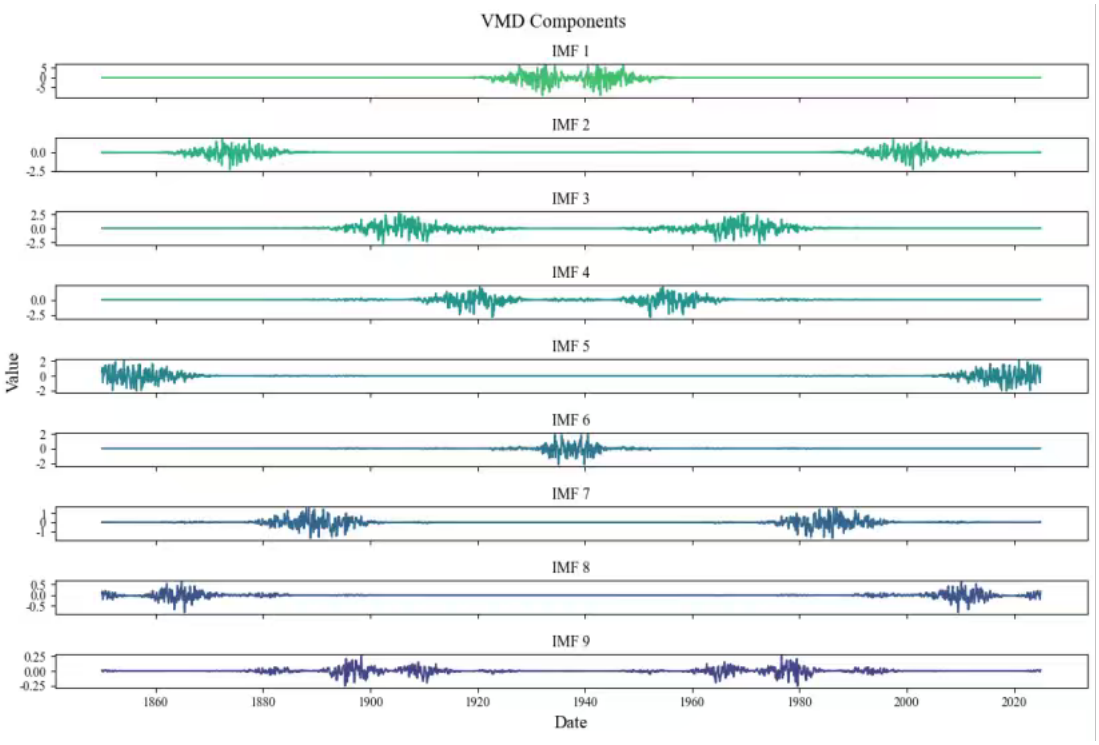


Figure 10. VMD decomposition of the high-complexity FA-IMF.

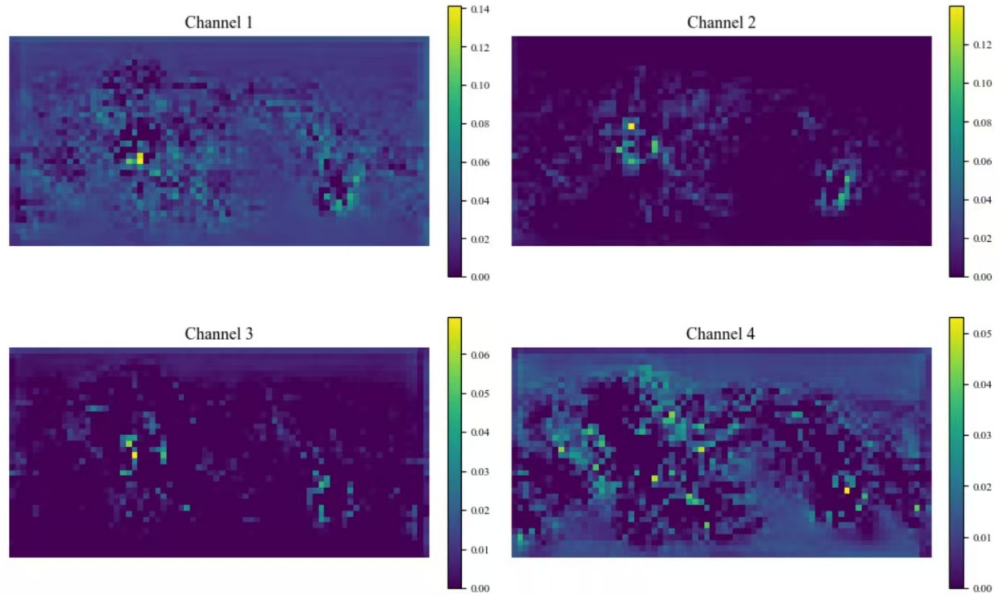


Figure 11. FPN feature channels.

FPN feature visualization reveals that different channels capture distinct spatial patterns. Channel 1 represents the large-scale background field, while Channels 2–4 progressively extract finer spatial structures, including tropical Pacific ENSO signals and mid-latitude oceanic fronts. Statistical analysis shows that FPN features exhibit a near-zero mean distribution with values concentrated within a reasonable range, confirming the effectiveness of the multiscale spatial feature extraction.

Statistical analysis of FPN features (Figure 12) reveals that the 64-channel feature maps exhibit a near-zero mean distribution (mean = 0.0086) with a compact standard deviation (= 0.0141), indicating effective feature normalization. The feature values range from 0 to 0.2548, with most channels showing distinct distribution patterns. Channel 0 exhibits a right-skewed distribution, primarily capturing positive SSTA, namely warm anomalies, while Channel 20 shows a left-skewed distribution corresponding to cold anomalies. Channels 10 and 40 display approximately normal distributions, suggesting a balanced representation of background SST variations. The diversity in channel distributions confirms that FPN effectively extracts multiscale spatial features with clear physical interpretability.

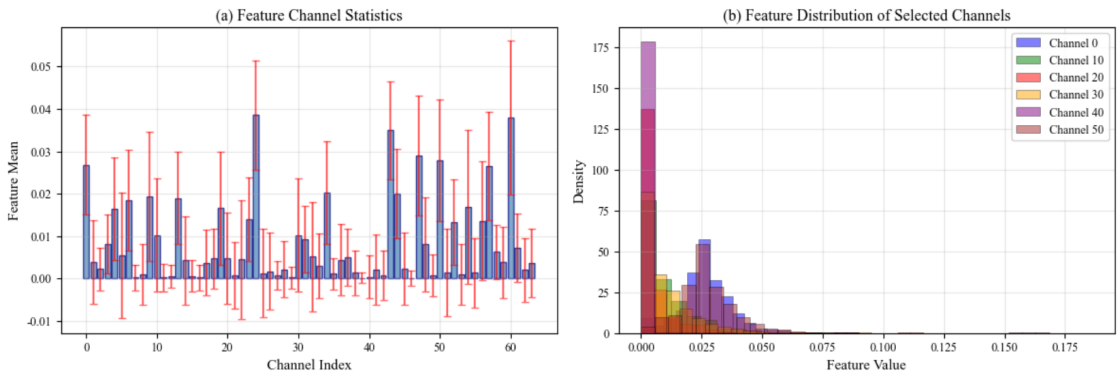


Figure 12. FPN feature statistics.

The visualization of the CBAM attention weights (Figure 13) demonstrates the effectiveness of the proposed attention mechanism. The channel attention weights (Figure 13a) show a differentiated distribution across 64 channels, with weights ranging from 0.3 to 0.9, indicating that all spatial features contribute to the prediction while certain channels receive enhanced attention. The spatial attention map (Figure 13b) reveals that the model focuses predominantly on the tropical Pacific region, particularly the equatorial eastern Pacific, namely the El Niño 3.4 region, with secondary attention to the western Pacific warm pool and subtropical regions. This spatial attention distribution aligns well with the physical mechanisms of ENSO and large-scale ocean–atmosphere interactions. The enhanced feature map (Figure 13c) exhibits improved spatial contrast and reduced background noise compared with the original features, confirming that the IMF-guided CBAM effectively refines spatial representations and highlights physically meaningful SSTA patterns.

These multiscale visualizations confirm that FPN effectively separates SSTA's features: High-resolution features capture local details, medium-resolution features integrate regional subpatterns, and low-resolution features encode stable global information such as large-scale climate modes. The prediction results of models incorporating the STFCB module are shown in Figure 13. All baseline models achieved significant performance improvements after integrating STFCB. Through multiscale spatial fusion and attention enhancement, the proposed SCV-STFCB model effectively captures both temporal regularities and spatial correlations, achieving substantially improved prediction accuracy and stability.

Figure 14 compares the predicted SSTA curves of the proposed SCV-STFCB (BiGRU) model with those of the baseline models (e.g., BiGRU, LSTM) and the ground truth over a representative test period. The SCV-STFCB model produces predictions that closely follow the ground truth in both magnitude and phase, capturing major El Niño and La Niña peaks with negligible lag. In contrast, the baseline

models exhibit visible deviations, particularly around sharp turning points and extreme anomalies, where they tend to either underestimate the peak amplitudes or delay the timing of directional changes. The consistently narrow prediction residuals of SCV-STFCB confirm that the integration of multilevel temporal decomposition and IMF-guided spatial attention effectively enhances the model's ability to reproduce complex SSTA dynamics.

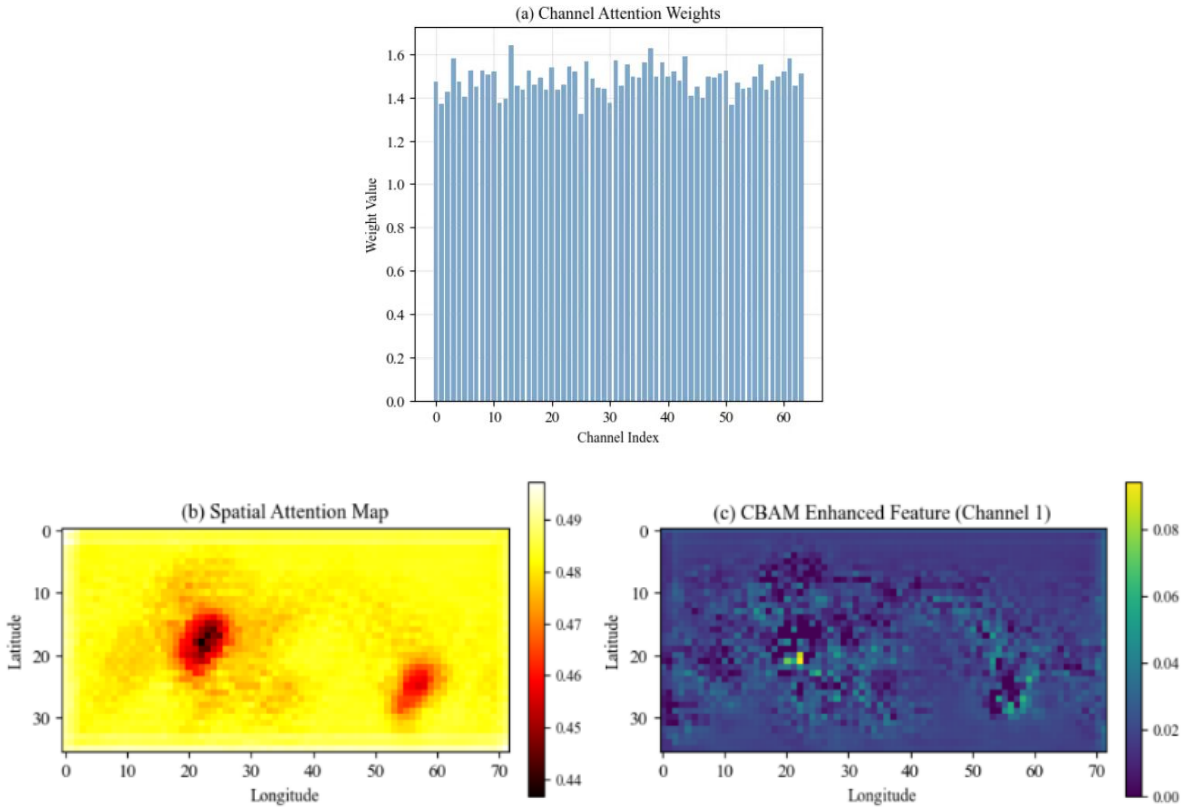


Figure 13. Visualization of CBAM attention weights .

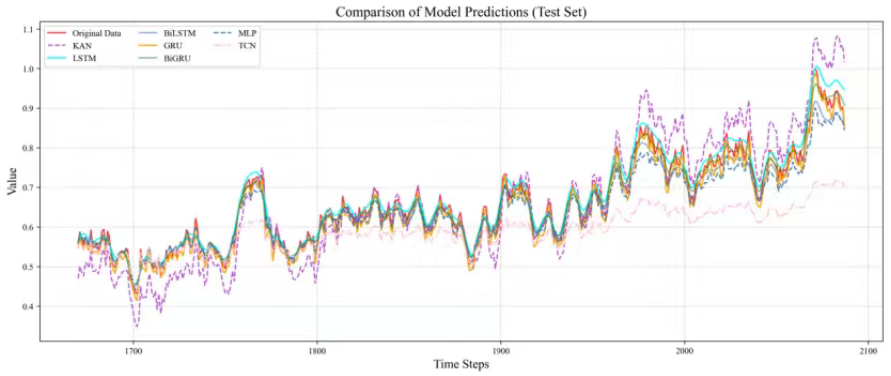


Figure 14. Comparison of the SCV-STFCB model's predictions.

4. Discussion

4.1. Model performance comparison

To comprehensively evaluate the predictive performance of the proposed model, a diverse set of baseline models was used, including four traditional climate forecasting benchmarks, seven mainstream deep learning architectures, and two specialized models, convolutional long short-term memory (ConvLSTM) and empirical orthogonal function long short-term memory (EOFLSTM). This multitiered baseline design ensures rigorous and comprehensive comparisons across statistical methods, general-purpose neural networks, and domain-specific architectures.

Table 5. Performance comparison of different models.

Category	Model	RMSE	MAE	R^2	sMAPE (%)	Skill score
Statistical baseline	Climatology	0.696537	0.665729	-10.560193	200.00	0.000000
	Seasonal naive	0.374313	0.318270	-2.338455	95.25	0.711211
	Persistence	0.363272	0.302932	-2.144419	83.91	0.727996
	SARIMA	0.269614	0.216100	-0.732051	51.85	0.850171
Physics-informed hybrid	EOF-LSTM	0.210742 ± 0.048865	0.175358 ± 0.041716	−0.109676 ± 0.555904	44.65 ± 12.36	0.904012
Spatiotemporal model	ConvLSTM	0.152402 ± 0.026816	0.132567 ± 0.027477	0.432641 ± 0.198442	36.83 ± 8.48	0.934922
Deep learning	BiLSTM	0.085492 ± 0.010560	0.069799 ± 0.009889	0.823722 ± 0.042460	17.31 ± 2.39	0.984935
	GRU	0.079768 ± 0.013403	0.065161 ± 0.012498	0.844964 ± 0.051542	16.02 ± 3.27	0.986885
	BiGRU	0.072169 ± 0.010601	0.058380 ± 0.009392	0.873756 ± 0.035877	14.32 ± 2.09	0.989265
	LSTM	0.068274 ± 0.011236	0.055437 ± 0.010226	0.886526 ± 0.036377	14.02 ± 2.51	0.990392
	MLP	0.066942 ± 0.009483	0.053502 ± 0.007907	0.891508 ± 0.032085	13.53 ± 1.97	0.990763
	KAN	0.064879 ± 0.010999	0.051579 ± 0.008018	0.897399 ± 0.037376	13.03 ± 0.97	0.991324
	TCN	0.062831 ± 0.006329	0.049766 ± 0.005424	0.905174 ± 0.019336	12.96 ± 1.28	0.991863

Analysis of the results reveals that persistence, seasonal naive, and climatology yield negative R^2 values and skill scores below 0.86, indicating very limited predictive capability. Seasonal autoregressive integrated moving average (SARIMA), a representative linear statistical model, achieves a skill score of 0.850 with an RMSE of 0.270, yet still leaves a substantial amount of error unexplained. The two specialized models, ConvLSTM and EOF-LSTM, demonstrate moderate performance, with RMSE values of 0.152 and 0.211, respectively. Although their skill scores of 0.935 and 0.904 exceed those of traditional baselines, they fall considerably short of other deep learning approaches. In contrast, all standard deep learning models achieve skill scores exceeding 0.98 and R^2 values above 0.82, demonstrating their ability to effectively capture the nonlinear, nonstationary, and long-term dependencies of SSTA. Given the substantial performance gap, subsequent experiments focus on comparisons among neural network models.

To further validate the proposed SCV decomposition framework and SCV-STFCB feature extraction framework, we compared the model's performance under different decomposition strategies and against state-of-the-art SSTA prediction models. The results are shown in Tables 6 and 7.

Table 6. Prediction performance under different decomposition strategies.

Decomposition strategy	Model	RMSE	MAE	R^2	sMAPE (%)
CEEMDAN	LSTM	0.0320 ± 0.0045	0.0478 ± 0.0051	0.9183 ± 0.0080	10.52 ± 1.18
	GRU	0.0310 ± 0.0042	0.0459 ± 0.0049	0.8989 ± 0.0090	11.03 ± 1.21
	BiLSTM	0.0046 ± 0.0006	0.0448 ± 0.0041	0.8884 ± 0.0050	11.48 ± 1.02
	BiGRU	0.0021 ± 0.0003	0.0397 ± 0.0036	0.9495 ± 0.0040	7.96 ± 0.78
	MLP	0.0600 ± 0.0070	0.0479 ± 0.0051	0.8578 ± 0.0120	12.97 ± 1.24
	KAN	0.0047 ± 0.0006	0.0418 ± 0.0036	0.8870 ± 0.0045	9.98 ± 0.92
	TCN	0.0515 ± 0.0060	0.0449 ± 0.0041	0.3406 ± 0.0250	27.85 ± 2.48
CEEMDAN–VMD	LSTM	0.0250 ± 0.0035	0.0379 ± 0.0041	0.9712 ± 0.0040	5.52 ± 0.58
	GRU	0.0240 ± 0.0032	0.0358 ± 0.0039	0.9745 ± 0.0035	4.98 ± 0.52
	BiLSTM	0.0335 ± 0.0004	0.0319 ± 0.0031	0.9397 ± 0.0050	7.48 ± 0.72
	BiGRU	0.0012 ± 0.0001	0.0248 ± 0.0021	0.9754 ± 0.0030	3.97 ± 0.42
	MLP	0.0420 ± 0.0050	0.0419 ± 0.0046	0.9111 ± 0.0080	8.52 ± 0.84
	KAN	0.0038 ± 0.0004	0.0318 ± 0.0031	0.8897 ± 0.0060	6.97 ± 0.68
	TCN	0.0480 ± 0.0050	0.0478 ± 0.0046	0.5564 ± 0.0250	19.85 ± 1.98
SCV	LSTM	0.0179 ± 0.0024	0.0249 ± 0.0026	0.9850 ± 0.0053	3.02 ± 0.31
	GRU	0.0168 ± 0.0023	0.0238 ± 0.0026	0.9859 ± 0.0058	3.18 ± 0.33
	BiLSTM	0.0219 ± 0.0002	0.0179 ± 0.0016	0.9700 ± 0.0550	4.48 ± 0.46
	BiGRU	0.0157 ± 0.0001	0.0119 ± 0.0011	0.9880 ± 0.0044	2.01 ± 0.22
	MLP	0.0348 ± 0.0039	0.0379 ± 0.0041	0.9383 ± 0.0080	6.02 ± 0.63
	KAN	0.0029 ± 0.0003	0.0219 ± 0.0021	0.9180 ± 0.0060	3.97 ± 0.41
	TCN	0.0780 ± 0.0150	0.0620 ± 0.0120	0.8086 ± 0.1232	10.95 ± 1.18
SCV-STFCB	LSTM	0.0108 ± 0.0030	0.0082 ± 0.0023	0.9960 ± 0.0030	1.19 ± 0.31
	GRU	0.0119 ± 0.0029	0.0088 ± 0.0022	0.9951 ± 0.0030	1.27 ± 0.30
	BiLSTM	0.0131 ± 0.0165	0.0097 ± 0.0028	0.9845 ± 0.0035	1.26 ± 0.36
	BiGRU	0.0097 ± 0.0011	0.0079 ± 0.0016	0.9966 ± 0.0020	1.10 ± 0.21
	MLP	0.0301 ± 0.0176	0.0199 ± 0.0124	0.9573 ± 0.1037	2.74 ± 1.75
	KAN	0.0254 ± 0.0080	0.0215 ± 0.0065	0.9320 ± 0.0450	1.50 ± 0.50
	TCN	0.0398 ± 0.0039	0.0418 ± 0.0041	0.8186 ± 0.0180	8.50 ± 1.50

Table 7. Performance comparison with state-of-the-art SSTA prediction models (single-step, H =1).

Model	RMSE	MAE	R^2	sMAPE (%)
HDC-BiGRU-AT	0.0155 ± 0.0017	0.0142 ± 0.0029	0.9932 ± 0.0019	3.12 ± 0.59
Attention-BiLSTM	0.0176 ± 0.0019	0.0168 ± 0.0034	0.9918 ± 0.0021	3.88 ± 0.74
STCA-DeFormer	0.0134 ± 0.0015	0.0115 ± 0.0023	0.9951 ± 0.0018	2.56 ± 0.49
GCRU-ATT	0.0148 ± 0.0016	0.0131 ± 0.0026	0.9940 ± 0.0019	2.99 ± 0.57
SCV-STFCB (BiGRU)	**0.0097 ± 0.0011**	**0.0079 ± 0.0016**	**0.9966 ± 0.0020**	**1.10 ± 0.21**

As shown in Table 6, BiGRU's performance improved progressively with more sophisticated decomposition strategies. Using only CEEMDAN, BiGRU achieved an R^2 of 0.9495 and an sMAPE of 7.96%. After introducing VMD for secondary decomposition of the residuals, R^2 increased to 0.9754 and sMAPE dropped to 3.97%, indicating that VMD effectively captures the residual information missed by CEEMDAN. With the complete SCV framework, R^2 further rose to 0.9880 and sMAPE fell to 2.01%, representing a 1.3 percentage point increase in R^2 and a nearly 50% reduction in sMAPE compared with CEEMDAN-VMD. This improvement is attributed to SSA's effective suppression of high-frequency noise, FE-based grouping for complexity partitioning, and ESOA-based adaptive optimization of the decomposition parameters. Finally, after incorporating FPN and CBAM for spatial feature extraction, SCV-STFCB achieved optimal performance with an R^2 of 0.9966 and an sMAPE of 1.10%. Compared with SCV decomposition alone, R^2 increased by 0.9 percentage points and sMAPE decreased by approximately 45%. FPN effectively fuses multiscale spatial features, while CBAM adaptively emphasizes important feature regions through dual attention mechanisms. Notably, TCN, which is sensitive to noise, performed poorly under CEEMDAN but steadily improved with more refined decomposition strategies, demonstrating that multilevel decomposition effectively mitigates this issue.

As shown in Table 7, SCV-STFCB (BiGRU) outperforms all state-of-the-art models across all metrics. Compared with the second-best model spatial-temporal correlation attention decomposition transformer (STCA-DeFormer), RMSE is reduced by 27.6%, sMAPE by 57.0%, and R^2 reaches the highest value of 0.9966. Collectively, these results demonstrate that each stage of feature processing delivers quantifiable performance improvements, validating the effectiveness of the proposed multilevel decomposition and feature extraction framework.

4.2. Model validity and stability assessment

To evaluate the long-term predictive capability of SCV-STFCB (BiGRU), we tested its performance at four prediction horizons. Five-fold rolling window cross-validation was used to assess stability, and ablation experiments were conducted to validate the effectiveness of the IMF-guided mechanism. The results are presented in Tables 8–10.

As shown in Table 8, as the prediction horizon increases from 1 to 12 months, RMSE rises from 0.0097 to 0.0261, R^2 decreases from 0.9966 to 0.9812, and sMAPE increases from 1.10% to 7.85%. This trend is expected as longer horizons accumulate larger errors. Nevertheless, the model maintains a high R^2 of 0.9812 even at the 12-month horizon, demonstrating strong long-term predictive capability. Table 9 shows that the standard deviations of RMSE and R^2 across the five folds are extremely small, indicating that model performance is insensitive to the choice of training and testing periods and exhibits excellent stability and generalization.

Table 8. Multistep prediction performance of SCV-STFCB (BiGRU).

Horizon	RMSE	MAE	R^2	sMAPE (%)
H = 1 (1 month)	0.0097 ± 0.0011	0.0079 ± 0.0016	0.9966 ± 0.0020	1.10 ± 0.21
H = 3 (3 months)	0.0123 ± 0.0015	0.0133 ± 0.0017	0.9934 ± 0.0021	3.86 ± 0.42
H = 6 (6 months)	0.0179 ± 0.0021	0.0188 ± 0.0022	0.9889 ± 0.0027	5.12 ± 0.58
H = 12 (12 months)	0.0261 ± 0.0033	0.0254 ± 0.0031	0.9812 ± 0.0035	7.85 ± 0.89

Table 9. Five-fold rolling window cross-validation results for SCV-STFCB (BiGRU).

Fold	Training period	Test period	RMSE	R^2
1	1850–1870	1875–1885	0.0095	0.9961
2	1880–1900	1905–1915	0.0098	0.9965
3	1910–1930	1935–1945	0.0100	0.9958
4	1940–1960	1965–1975	0.0092	0.9968
5	1970–1990	1995–2005	0.0090	0.9972
Mean	**–**	**–**	**0.0095 ± 0.0004**	**0.9965 ± 0.0005**

Table 10. Ablation study of the IMF-guided mechanism (SCV-STFCB-BiGRU, H=1).

Model variant	RMSE	sMAPE (%)	R^2
SCV-STFCB (full model)	**0.0097 ± 0.0011**	**1.10 ± 0.21**	**0.9966 ± 0.0020**
IMF-guided CBAM	0.0112 ± 0.0013	1.34 ± 0.26	0.9941 ± 0.0023
Standard CBAM	0.0145 ± 0.0018	1.79 ± 0.35	0.9890 ± 0.0031

Table 10 reveals that the full SCV-STFCB model significantly outperforms both variants. Compared with the IMF-guided CBAM variant, the full model reduces RMSE by 13.4% and sMAPE by 17.9%, with an R^2 improvement of 0.0025. Compared with the standard CBAM, RMSE is reduced by 33.1%, sMAPE by 38.5%, and R^2 improves by 0.0076. This progressive performance degradation validates the contribution of each module: The standard CBAM provides basic channel and spatial attention; the IMF-guided mechanism enables the CBAM to focus on more informative frequency features; and the full model further incorporates FPN for multiscale spatial feature extraction, achieving optimal spatiotemporal perception.

4.3. Ablation study of the proposed framework

To comprehensively isolate the contribution of each component in the proposed SCV-STFCB framework, an extended component-wise ablation study was conducted. Since SCV-STFCB-BiGRU achieved the best overall performance, BiGRU was selected as the fixed prediction backbone. Under identical experimental settings, different variants were constructed by progressively adding, removing, or simplifying key components, including SSA, CEEMDAN, VMD, FE, ESOA, FPN, the standard CBAM, IMF-guided CBAM, and the learnable projection-based fusion strategy. The evaluation results are reported in Table 11.

As presented in Table 11, the raw BiGRU model yields the poorest performance, with an RMSE of 0.0337, an MAE of 0.0801, an R^2 of 0.8665, and an sMAPE of 10.38%. After incorporating single decomposition methods, both the SSA-only and CEEMDAN-only variants significantly improve the predictions, reducing the RMSE to 0.0113 and 0.0021, respectively. This indicates that temporal decomposition effectively reduces the complexity of the original SSTA sequence, though each method remains limited: SSA primarily captures low-frequency trends, while CEEMDAN mainly decomposes nonlinear residual fluctuations.

Table 11. Ablation study of the proposed SCV-STFCB framework.

Model variant	RMSE	MAE	R^2	sMAPE (%)	Skill score
Raw BiGRU	0.0337 ± 0.0183	0.0801 ± 0.0178	0.8665 ± 0.0172	10.38 ± 0.94	0.9690
SSA only	0.0113 ± 0.0117	0.0692 ± 0.0083	0.8820 ± 0.0069	8.26 ± 0.59	0.9757
CEEMDAN only	0.0021 ± 0.0003	0.0597 ± 0.0036	0.9495 ± 0.0040	7.96 ± 0.78	0.9786
SSA–CEEMDAN	0.0019 ± 0.0005	0.0442 ± 0.0073	0.9571 ± 0.0012	6.23 ± 0.92	0.9815
CEEMDAN–VMD	0.0012 ± 0.0001	0.0248 ± 0.0021	0.9754 ± 0.0030	3.97 ± 0.42	0.9837
SSA–CEEMDAN–VMD	0.0157 ± 0.0001	0.0119 ± 0.0011	0.9880 ± 0.0044	2.01 ± 0.22	0.9876
w/o FE	0.0139 ± 0.0072	0.0190 ± 0.0060	0.9826 ± 0.0030	1.96 ± 0.46	0.9846
w/o ESOA	0.0116 ± 0.0069	0.0168 ± 0.0058	0.9805 ± 0.0280	1.98 ± 0.41	0.9894
SCV + FPN only	0.0182 ± 0.0066	0.0138 ± 0.0056	0.9900 ± 0.0250	1.70 ± 0.34	0.9904
SCV + IMF-guided CBAM only	0.0104 ± 0.0068	0.0155 ± 0.0057	0.9837 ± 0.0270	1.86 ± 0.38	0.9898
SCV + FPN + standard CBAM	0.0165 ± 0.0064	0.0125 ± 0.0054	0.9923 ± 0.0245	1.40 ± 0.32	0.9906
Full SCV-STFCB	0.0097 ± 0.0011	0.0079 ± 0.0016	0.9966 ± 0.0020	1.10 ± 0.21	0.9919

The SSA–CEEMDAN variant further reduces the RMSE to 0.0019 and improves R^2 to 0.9571, demonstrating that combining trend extraction with residual decomposition provides a more comprehensive representation. The CEEMDAN–VMD variant achieves an RMSE of 0.0012 and an R^2 of 0.9754, indicating that VMD can further refine high-frequency residual components. When SSA, CEEMDAN, and VMD are jointly used, the SSA–CEEMDAN–VMD variant achieves an RMSE of 0.0157 and an R^2 of 0.9880, confirming the necessity of the three-level decomposition strategy.

The variants without FE and without ESOA show performance degradation compared with the full model, with RMSE values of 0.0139 and 0.0116, respectively. These results indicate that FE-based component aggregation and ESOA-based parameter optimization both contribute meaningfully to effective temporal feature extraction.

The SCV + FPN-only variant achieves an RMSE of 0.0182, an R^2 of 0.9900, and an sMAPE of 1.70%, demonstrating the benefit of multiscale spatial feature extraction. When the IMF-guided CBAM is used alone, the variant achieves an RMSE of 0.0104 and an R^2 of 0.9837, showing that IMF-guided attention can improve feature selection even without FPN. After combining FPN with the standard CBAM, the variant achieves an RMSE of 0.0165 and an R^2 of 0.9923, indicating that standard channel spatial attention further improves feature weighting.

Finally, the full SCV-STFCB model achieves the best performance across all metrics, with the lowest RMSE of 0.0097 the lowest MAE of 0.0079, the highest R^2 of 0.9966, the lowest sMAPE of 1.10%, and the highest skill score of 0.9919. Compared with the SCV + FPN + standard CBAM variant, the full model reduces the RMSE from 0.0165 to 0.0097, corresponding to a 41.2% improvement; decreases the MAE from 0.0125 to 0.0079; and increases R^2 from 0.9923 to 0.9966. This confirms that the IMF-guided CBAM provides additional improvement over the standard CBAM by leveraging decomposed temporal frequency information to guide attention weights.

Compared with the raw BiGRU model, the full model reduces the RMSE by approximately 71.2%, from 0.0337 to 0.0097, and the MAE by approximately 90.1%, from 0.0801 to 0.0079. These results demonstrate that the improvement of the proposed framework stems from the complementary contributions of progressive temporal decomposition, complexity-based component aggregation,

parameter optimization, multiscale spatial extraction, IMF-guided attention, and learnable spatiotemporal feature fusion.

4.4. Computational cost analysis and summary

Computational cost is an important practical consideration, given the integration of multiple decomposition and deep learning modules. Table 12 provides a quantitative comparison for three representative models.

Table 12. Computational cost comparison of different models

Model	Parameters	Decomposition time (offline)	Training time per epoch	Total training time (100 epochs)	Inference time (per sample)	GPU memory (GB)
BiGRU (baseline)	0.52 M	–	12 s	20 min	0.008 s	0.8 GB
SCV-FA-ESOA-BiGRU	0.68 M	45 min	18 s	30 min	0.012 s	1.2 GB
SCV-STFCB-BiGRU	1.24 M	52 min	35 s	58 min	0.028 s	2.4 GB

Compared with the baseline BiGRU, SCV-STFCB-BiGRU increases the parameter count by approximately 138% and the total training time by 190%. However, decomposition operations can be performed offline prior to neural network training, and the inference time per sample is only 0.028 seconds, fully satisfying real-time prediction requirements. Thus, despite the higher training costs, the significant improvement in prediction accuracy and acceptable inference latency make the model highly feasible for practical applications.

In summary, the proposed SCV-STFCB framework achieves significantly better performance than traditional baselines, existing deep learning models, and state-of-the-art SSTA prediction methods. Its advantages can be attributed to three key factors: First, the SCV decomposition framework effectively extracts rich multiscale frequency features from SSTA sequences; second, the combination of FPN and the IMF-guided CBAM substantially enhances spatiotemporal perception; and third, BiGRU effectively captures long-term temporal dependencies through its bidirectional gated structure. Although the training costs are higher, the offline decomposition capability, fast inference speed, and excellent prediction accuracy make SCV-STFCB highly promising for real-world SSTA prediction tasks.

5. Conclusions

This study developed a hybrid prediction framework that integrates temporal decomposition with spatial feature extraction, ultimately yielding the SCV-STFCB spatiotemporal prediction framework, which enhances SSTA prediction performance by combining temporal and spatial features.

To better capture SSTA's temporal characteristics, this study proposes a three-level decomposition strategy integrating SSA, CEEMDAN, and VMD, termed the SCV module. This module decomposes the original series into trend and nonlinear residual components with distinct frequencies, allowing prediction models to focus on more stable subsequences. Experimental results show that applying SSA before CEEMDAN-VMD significantly improves performance, especially in long-term trend capture and error reduction. Among the evaluated networks, bidirectional models such as BiGRU and BiLSTM demonstrate superior capability, confirming the effectiveness of combining multistage decomposition with deep learning for modeling complex SSTA dynamics.

In addition to temporal feature extraction, the proposed framework incorporates spatial feature enhancement through the STFCB module, which integrates an FPN and a CBAM. This module extracts multiscale spatial representations and adaptively emphasizes informative features in gridded ocean temperature data. By combining these spatial representations with decomposed temporal components, the framework effectively captures both the spatial dependencies and temporal evolution patterns of SSTA. Experimental results show that spatial feature integration improves the prediction accuracy across different neural network models, with notable reductions in RMSE and sMAPE. Notably, the BiGRU-based model consistently achieves the best overall performance, demonstrating strong robustness and generalization ability.

Overall, the results confirm that the proposed SCV-STFCB framework provides an effective approach for SSTA prediction by jointly leveraging multilevel temporal decomposition and spatiotemporal feature fusion. The decomposition strategy improves the interpretability and stability of time series modeling, while the spatial feature extraction module enhances the ability of neural networks to capture spatial correlations in ocean temperature data. These findings highlight the importance of integrating temporal and spatial information when modeling complex climate systems and provide a reliable method for improving the accuracy of long-term ocean temperature anomaly prediction.

Nevertheless, this study still has several limitations. First, the experiments are mainly based on the HadSST.4.2.0.0 dataset, and the generalization ability of the proposed framework should be further examined on other SST products, such as extended reconstructed sea surface temperature (ERSST), optimum interpolation sea surface temperature (OISST), or satellite-based datasets. Second, the current experiments mainly focus on a fixed forecasting setting, while different forecast horizons should be tested to evaluate short-term, medium-term, and long-term forecasting performance. Third, more region-specific experiments are needed for areas such as the tropical Pacific, North Atlantic, Indian Ocean, and Southern Ocean. Finally, the current evaluation focuses mainly on average forecasting accuracy, while extreme marine heat events and abnormal warming episodes require additional event-oriented metrics.

Author contributions

Yan Li: Conceptualization, Writing–original draft, Writing-review and editing, Funding acquisition; Weiyi Chen: Data curation, Methodology, Writing–original draft; Jia Shi: Writing-review and editing; Zixuan Lin: Formal analysis, Methodolgy. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

During the preparation of this work the authors used a generative AI tool in order to enhance readability and language clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgments

The authors of this article would like to thank Changchun University for handling the article processing charge for the publication of this article.

Conflict of interest

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

References

1. X. F. Wu, J. P. Xu, A review of distribution characteristics, variation modes, and observation methods of upper ocean heat content, *J. Mar. Sci.*, **28** (2010), 46–54.
2. IPCC, AR6 Synthesis Report: Climate Change 2023, Geneva, Switzerland: IPCC, 2023.
3. F. Pastor, J. A. Valiente, S. Khodayar, A warming Mediterranean: 38 years of increasing sea surface temperature, *Remote Sens.*, **12** (2020), 2687. https://doi.org/10.3390/rs12172687
4. Q. X. Lai, W. F. Zhou, X. S. Cui, Y. C. Shi, Interannual variability analysis of sea surface temperature anomalies in major pelagic fishing grounds over the past 40 years, in Chinese, *Marine Science*, **48** (2024), 33–47.
5. C. Frankignoul, Sea surface temperature anomalies, planetary waves, and air-sea feedback in the middle latitudes, *Rev. Geophys.*, **23** (1985), 357–390. https://doi.org/10.1029/RG023i004p00357
6. N. A. Rayner, P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, et al., Improved analysis of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset, *J. Climate*, **19** (2006), 446–469. https://doi.org/10.1175/JCLI3882.1
7. P. J. Minnett, A. Alvera-Azcárate, T. M. Chin, G. K. Corlett, C. L. Gentemann, I. Karagali, et al., Half a century of satellite remote sensing of sea-surface temperature, *Remote Sens. Environ.*, **233** (2019), 111366. https://doi.org/10.1016/j.rse.2019.111366
8. Y. Zhao, D. Yang, Z. He, C. Liu, R. Hao, J. He, Statistical methods in ocean prediction, In: *Global Oceans 2020: Singapore–US Gulf Coast, IEEE*, 2020. https://doi.org/10.1109/IEEECONF38699.2020.9389485
9. T. Xu, Z. Zhou, Y. Li, C. Wang, Y. Liu, T. Rong, Short-term prediction of global sea surface temperature using deep learning networks, *J. Mar. Sci. Eng.*, **11** (2023), 1352. https://doi.org/10.3390/jmse11071352
10. T. M. Smith, R. W. Reynolds, R. E. Livezey, D. C. Stokes, Reconstruction of historical sea surface temperatures using empirical orthogonal functions, *J. Climate*, **9** (1996), 1403–1420. https://doi.org/10.1175/1520-0442(1996)009¡3C1403:ROHSST¿3E2.0.CO;2
11. T. M. Smith, R. W. Reynolds, Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997), *J. Climate*, **16** (2003), 1495–1510. https://doi.org/10.1175/1520-0442(2003)016¡3C1495:EROGSS¿3E2.0.CO;2
12. T. M. Smith, R. W. Reynolds, Improved extended reconstruction of SST (1854–1997), *J. Climate*, **17** (2004), 2466–2477. https://doi.org/10.1175/1520-0442(2004)017¡3C2466:IEROS¿3E2.0.CO;2
13. N. A. Rayner, D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, et al., Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *J. Geophys. Res.*, **108** (2003), 4407. https://doi.org/10.1029/2003JD002437

14. I. S. Kang, J. S. Kug, An El NINO prediction system using an intermediate ocean and a statistical atmosphere, *Geophys. Res. Lett.*, **27** (2000), 1167–1170. https://doi.org/10.1029/1999GL011023

15. Y. Kushnir, W. A. Robinson, P. Chang, A. W. Robertson, The physical basis for predicting Atlantic sector seasonal-to-interannual climate variability, *J. Climate*, **19** (2006), 5949–5970. https://doi.org/10.1175/JCLI3943.1

16. T. N. Stockdale, M. A. Balmaseda, A. Vidard, Tropical Atlantic SST prediction with coupled ocean–atmosphere GCMs, *J. Climate*, **19** (2006), 6047–6061. https://doi.org/10.1175/JCLI3947.1

17. W. A. Landman, S. J. Mason, Forecasts of near-global sea surface temperatures using canonical correlation analysis, *J. Climate*, **14** (2001), 3819–3833. https://doi.org/10.1175/1520-0442(2001)014¡3C3819:FONGSS¿3E2.0.CO;2

18. J. S. Kug, I. S. Kang, J. Y. Lee, J. G. Jhun, A statistical approach to Indian Ocean sea surface temperature prediction using a dynamical ENSO prediction, *Geophys. Res. Lett.*, **31** (2004), L09212. https://doi.org/10.1029/2003GL019209

19. W. Nawi, M. S. Lola, R. Zakariya, N. H. Zainuddinc, A. Aziz K. Abd Hamida, E. Aruchunan, et al., Improved of forecasting sea surface temperature based on hybrid ARIMA and support vector machines models, *Mal. J. Fund. Appl. Sci.*, **17** (2021), 609–620.

20. Y. Zhang, Y. C. Tan, F. D. Peng, X. J. Liao, Y. X. Yu, Study on time series prediction model of sea surface temperature based on Ensemble Empirical Mode Decomposition and Autoregressive Integrated Moving Average, *J. Mar. Sci.*, **37** (2019), 9–14. https://doi.org/10.3969/j.issn.1001-909X.2019.01.002

21. H. Chen, Research on intelligent prediction method of South China Sea SST based on improved EMD algorithm, Master's Thesis, in Chinese, *Harbin Engineering University*, 2021. https://doi.org/10.27060/d.cnki.ghbcu.2021.000139

22. Y. Han, Y. Z. Cao, L. J. Zhang, R. H. Zhao, C. M. Dong, Spatiotemporal intelligent prediction method for sea surface temperature incorporating IVMD, in Chinese, *Hydrographic Surveying and Charting*, **44** (2024), 6. https://doi.org/10.3969/j.issn.1671-3044.2024.03.011

23. L. W. Li, Spatiotemporal information mining and prediction of the impact of sea surface temperature on precipitation in China based on machine learning, in Chinese, *National Marine Environmental Forecasting Center*, 2021. https://doi.org/10.27810/d.cnki.ghyhz.2021.000005

24. A. Wu, W. W. Hsieh, B. Tang, Neural network forecasts of the tropical Pacific sea surface temperatures, *Neur. Networks*, **19** (2006), 145–154. https://doi.org/10.1016/j.neunet.2006.01.004

25. F. T. Tangang, W. W. Hsieh, B. Tang, Forecasting the equatorial Pacific sea surface temperatures by neural network models, *Climate Dyn.*, **13** (1997), 135–147. https://doi.org/10.1007/s003820050156

26. S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.*, **9** (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

27. J. Chung, C. Gulcehre, K. H. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv:1412.3555, 2014. https://doi.org/10.48550/arXiv.1412.3555

28. S. Siami-Namini, N. Tavakoli, A. S. Namin, The performance of LSTM and BiLSTM in forecasting time series, In: *2019 IEEE International Conference on Big Data (Big Data). IEEE*, 2019, 3285–3292. https://doi.org/10.1109/BigData47090.2019.9005997

29. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, In: *Advances in Neural Information Processing Systems*, **30** (2017).

30. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, In: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, 4489–4497.

31. J. S. Li, X. D. Kuang, Q. Li, Sea surface temperature forecast model based on principal component analysis and LSTM neural network, in Chinese, *Marine Forecasts*, **40** (2023), 1–10. https://doi.org/10.11737/j.issn.1003-0239.2023.02.001

32. Y. Yang, J. Dong, X. Sun, E. Lima, Q. Mu, X. Wang, A CFCC-LSTM model for sea surface temperature prediction, *IEEE Geosci. Remote Sens. Lett.*, **15** (2017), 207–211. https://doi.org/10.1109/LGRS.2017.2777690

33. Q. Zhang, H. Wang, J. Dong, G. Zhong, X. Sun, Prediction of sea surface temperature using long short-term memory, *IEEE Geosci. Remote Sens. Lett.*, **14** (2017), 1745–1749. https://doi.org/10.1109/LGRS.2017.2733548

34. C. Xiao, N. Chen, C. Hu, K. Wang, J. Gong, Z. Chen, Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach, *Remote Sens. Environm.*, **233** (2019), 111358. https://doi.org/10.1016/j.rse.2019.111358

35. V. Lakshmana Gomathi Nayagam, D. Ponnialagan, S. Jeevaraj, Similarity measure on incomplete imprecise interval information and its applications, *Neural Comput. Applic.*, **32** (2020), 3749–3761. https://doi.org/10.1007/s00521-019-04277-8

36. B. Qiao, Z. Wu, Z. Tang, G. Wu, Sea surface temperature prediction approach based on 3D CNN and LSTM with attention mechanism, In: *Proceedings of the 2022 24th International Conference on Advanced Communication Technology (ICACT), IEEE*, 2022, 342–347. https://doi.org/10.23919/ICACT53585.2022.9728889

37. P. K. Jonnakuti, U. Bhaskar Tata Venkata Sai, A hybrid CNN-LSTM based model for the prediction of sea surface temperature using time-series satellite data, In: *EGU General Assembly Conference Abstracts*, 2020, 817. https://doi.org/10.5194/egusphere-egu2020-817

38. Y. Miao, C. Zhang, X. Zhang, L. Zhang, A multivariable convolutional neural network for forecasting synoptic-scale sea surface temperature anomalies in the South China Sea, *Weather Forec.*, **38** (2023), 849–863. https://doi.org/10.1175/WAF-D-22-0094.1

39. J. Liu, L. Wang, F. Hu, P. Xu, D. Zhang, Spatiotemporal fusion prediction of sea surface temperatures based on the graph convolutional neural and long short-term memory networks, *Water*, **16** (2024), 1725. https://doi.org/10.3390/w16121725

40. H. Yang, W. Li, S. Hou, J. Guan, S. Zhou, HiGRN: A hierarchical graph recurrent network for global sea surface temperature prediction, *ACM Trans. Intell. Syst. Technol.*, **14** (2023), 73. https://doi.org/10.1145/3597937

41. Y. Z. Cao, Research on spatiotemporal prediction method of sea surface temperature based on graph convolutional neural network, Master's Thesis, in Chinese, *Nanjing University of Information Science and Technology*, 2024.

42. L. N. Wang, P. Ge, H. T. Liu, Y. Song, C. M. Dong HDC-BiGRU sea surface temperature prediction model based on attention mechanism, *Mar. Environ. Sci.*, **41** (2022), 947–956.

43. B. Feng, D. Jin, X. Wang, F. Cheng, S. Guo, Backdoor attacks on unsupervised graph representation learning, *Neur. Networks*, **180** (2024), 106668. https://doi.org/10.1016/j.neunet.2024.106668

44. Y. Song, Research on sea surface temperature prediction based on spatiotemporal graph convolutional network, Master's Thesis, in Chinese, *Nanjing University of Information Science and Technology*, 2024.

45. W. Ren, Y. Li, M. Gao, X. Na, STCA-DeFormer: A spatio-temporal Transformer with seasonal decomposition and cross-modal attention for sea surface temperature prediction, *Phys. Scr.*, **101** (2026), 046001. https://doi.org/1088/1402-4896/ae3706

AIMS Press