remote sensing

MDPI

Article

Spatial Downscaling of Sea Surface Temperature Using Diffusion Model

Shuo Wang †, Xiaoyan Li †, Xueming Zhu *, Jiandong Li and Shaojing Guo

Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Marine Sciences, Sun Yat-sen University, Zhuhai 519082, China; wangsh557@mail2.sysu.edu.cn (S.W.); lixiaoyan@sml-zhuhai.cn (X.L.); lijd36@mail2.sysu.edu.cn (J.L.); guoshj9@mail2.sysu.edu.cn (S.G.)

* Correspondence: zhuxueming@sml-zhuhai.cn

† These authors contributed equally to this work and should be regarded as co-first authors.

Abstract: In recent years, advancements in high-resolution digital twin platforms or artificial intelligence marine forecasting have led to the increased requirements of high-resolution oceanic data. However, existing sea surface temperature (SST) products from observations often fail to meet researchers' resolution requirements. Deep learning models serve as practical techniques for improving the spatial resolution of SST data. In particular, diffusion models (DMs) have attracted widespread attention due to their ability to generate more vivid and realistic results than other neural networks. Despite DMs' potential, their application in SST spatial downscaling remains largely unexplored. Hence we propose a novel DM-based spatial downscaling model, called DIFFDS, designed to obtain a high-resolution version of the input SST and to restore most of the meso scale processes. Experimental results indicate that DIFFDS is more effective and accurate than baseline neural networks, its downscaled high-resolution SST data are also visually comparable to the ground truth. The DIFFDS achieves an average root-mean-square error of 0.1074 °C and a peak signal-to-noise ratio of 50.48 dB in the 4× scale downscaling task, which shows its accuracy.

Keywords: spatial downscaling; diffusion model; sea surface temperature; deep learning

Citation: Wang, S.; Li, X.; Zhu, X.; Li, J.; Guo, S. Spatial Downscaling of Sea Surface Temperature Using Diffusion Model. *Remote Sens.* **2024**, *16*, 3843. https://doi.org/10.3390/rs16203843

Academic Editor: Andrea Storto

Received: 30 August 2024
Revised: 14 October 2024
Accepted: 14 October 2024
Published: 16 October 2024

1. Introduction

The sea surface temperature (SST) is a crucial climate variable that contributes to Earth's climate [1]. SST influences marine ecosystems, ocean–atmosphere interactions, and oceanic currents. Recent advances in high-resolution digital twin platforms or artificial intelligence marine forecasting, such as Earth-2 (with kilometer-scale resolution), have led to an increased demand for high-resolution oceanic data. High-resolution SST data can reveal more meso- or small-scale dynamic processes, enabling neural networks to learn more complex patterns.

However, due to the limitations of current observation technology, the resolution of existing satellite remote sensing SST products often fails to meet researchers' needs. This severely restricts the potential applications of SST data in various fields, like deep learning-based oceanographic models.

To mitigate this issue, oceanographers have begun to use spatial downscaling techniques to obtain higher-resolution SST datasets. By establishing mapping relationships between low- and high-resolution data, spatial downscaling can generate high-resolution versions of input low-resolution SST. The downscaled high-resolution outcomes can reveal more detailed marine dynamic features, providing a higher spatial resolution for subsequent applications.

Downscaling techniques can be categorized into dynamic downscaling, statistical downscaling, and deep learning-based downscaling methods. Dynamic downscaling is conducted by nesting regional models into low-resolution global models to produce high-resolution information. For instance, Huang et al. [2] used a variable-resolution option

within the community earth system model to simulate California's climate, demonstrating competitive utility for studying high-resolution regional climatology compared to the weather research and forecast model. Dynamic downscaling can produce reliable results because it uses physical equations to describe dynamic processes, but it often entails high spatial-temporal complexity. In addition to dynamic downscaling, statistical downscaling methods are also widely employed to establish empirical relationships between large-scale variables and local-scale parameters to produce high-resolution data [3]. For instance, Jorge et al. [4] proposed a weather-type downscaling method for multivariate ocean wave climate based on a statistical downscaling framework. Statistical downscaling faces challenges in accurately establishing statistical relationships in areas with complex terrain or unique climate zones, and the quality of the input data affects the precision of the result.

Deep learning has shown impressive prospects in various tasks within marine science, including downscaling, prediction, and reconstruction of oceanic elements. Deep learning-based downscaling originates primarily from super-resolution (SR) technology in computer vision. Dong et al. [5] proposed the first super-resolution convolutional neural network (SRCNN). Subsequently, many CNN-based SR models were created, and their performance improved further [6–10]. The emergence of SR generative adversarial models such as enhanced SR generative adversarial network (ESRGAN) [11], and SR transformer models [12,13], also represent significant advances in SR.

Deep learning methods can generate accurate high-resolution data. They can also automatically learn feature representations from oceanic data, allowing for more effective feature extraction. These characteristics have prompted many oceanographers to explore deep-learning downscaling methods. Some researchers use interpolation algorithms, such as bicubic interpolation or nearest neighbor interpolation, to obtain a low-resolution SST from a high-resolution SST. They then employ these low-resolution SSTs as the input and the original high-resolution SST as the target to train neural networks for the spatial downscaling tasks. For instance, Aurelien and Ronan [14] utilized bicubic interpolation to generate low-resolution input from the operational sea surface temperature and sea ice analysis (OSTIA) dataset and then employed SRCNN to generate high-resolution targets. Similarly, Khoo et al. developed an SN-ESRGAN neural network to downscale low-resolution SST into high-resolution ones [15], wherein the nearest neighbor algorithm was used for generating input data from OSTIA SST. These approaches highlight the efficacy of deep learning-based spatial downscaling techniques in addressing SST downscaling challenges. In other scenarios, high-resolution infrared SST data and low-resolution microwave SST are used to train deep learning models. Izumi et al. trained a CNN-based network with 125 km resolution input and 25 km resolution ground truth, achieving high-quality results [16]. Zou et al. designed a transformer-based model to obtain a resolution output of 0.02°, using the 0.25° advanced microwave scanning radiometer 2 SST as input [17].

Recently, diffusion models (DM) have gained significant attention along with the rise of text-to-image generation models such as Imagen [18] and DALL-E2 [19]. Unlike other deep neural networks, DMs excel at producing more vivid samples and circumventing issues like mode collapse, which can be seen in GANs. These advantages have led to their broad application across various computer vision domains. In SR, diffusion model-based approaches, exemplified by works such as [20–22] have achieved remarkable results. Nevertheless, studies and practical applications focusing on DM-based SST downscaling are noticeably scarce. To explore whether the generative capabilities of DM can be harnessed to restore missing details and processes in low-resolution SSTs, we propose a novel spatial downscaling method DIFFDS based on diffusion model for image restoration (DIFFIR) [23]. In comparison to the original DIFFIR, DIFFDS redesigns the transformer block by introducing cross-attention and channel-attention mechanisms. This refinement results in fewer abnormal textures and more mesoscale details, making DIFFDS more suitable for SST spatial downscaling. We conducted 4× scale downscaling experiments to reveal the superior performance of DIFFDS over several existing methods, offering a fresh perspective for the SST downscaling field.

Our contributions can be summarized as follows:

1. We extended the application of DM to address SST spatial downscaling problems. The proposed DIFFDS method leveraged the robust distribution fitting and generation capabilities of DM to reconstruct high-resolution SSTs. Experimental results demonstrate its effectiveness.
2. To ensure greater consistency between the downscaling results and the original high-resolution SSTs and to mitigate incorrect texture anomalies in the results, we restructured the transformer block in DIRformer. This enhancement allows DIFFDS to fully consider the underlying structure in low-resolution data, thereby resulting in a more reasonable reconstruction of high-resolution SST contents.
3. DIFFDS outperforms the commonly used CNN, GAN, and regression methods on most evaluation metrics and closely approximates the visual fidelity of high-resolution ground truth. This substantiates its superiority over other models.

2. Materials and Methods

2.1. Study Area and Data

2.1.1. Study Area

As shown in Figure 1, the study sea area extends from 6°N to 22°N, 107°E to 123°E. It encompasses the South China Sea, the Luzon Strait, and the Sulu Sea. This domain experiences a prevailing tropical maritime monsoon climate characterized by warm temperatures, seasonal monsoons, and significant rainfall. These conditions induce complex SST distributions and multi-scale dynamical processes, such as upwelling, mesoscale eddies, and oceanic fronts.

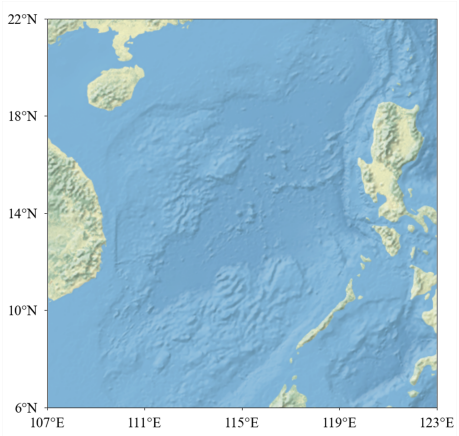

Figure 1. The study area used in this paper.

2.1.2. SST Data

The SST data employed in this study comes from operational sea surface temperature and sea ice analysis reprocessed (OSTIA-REP) [24–26] SST dataset, which is a group for high-resolution sea surface temperature (GHRSST) generated by using optimal interpolation (OI) on a global 0.05° degree grid. As a sister product to the near real-time counterpart (OSTIA-NRT), the OSTIA-REP distinguishes itself by assimilating satellite data from more than 25 distinct SST sensors, along with in situ observations sourced from drifting and moored buoys.

The original data resolution in the study region is 320 × 320 pixels (0.05°). To ensure a more acceptable training speed, we resized the original SST data by using the nearest neighbor downsampling algorithm to obtain a lower-resolution version of the data. The downsampled 48 × 48 pixels ($\frac{1}{3}$°) SST data and 192×192 pixels ($\frac{1}{12}$°) SST data are then considered as the low-resolution input and the corresponding high-resolution target in the 4× downscaling experiments.

In order to distinguish the land and sea points, the SST values over the land points are first set to 0. Then, since all SST values in the dataset are less than 35, the original SST values are normalized to SST′ within the data range $[-1, 1)$, according to Equation (1). This normalization preprocessing is conducted to ensure fairness in the comparison between the proposed DIFFDS method and other algorithms in the experimental results:

$$SST' = 2 \times (\frac{SST}{35}) - 1 \quad (1)$$

The period of the experiment dataset covers from 1990 to 2021. Data from 1990 to 2019 are used as the training set, data within 2020 are the validation set, and the data from March 2021 to February 2022 serve as the test set.

2.2. Diffusion Model

Generative models, such as GANs and variational autoencoders (VAEs), are commonly used to produce highly realistic samples but inherently come with limitations. GANs, for instance, can encounter challenges in training stability and sampling diversity unless carefully designed optimization strategies are employed. In addition, these GAN methods may easily suffer from mode collapse [27]. This phenomenon often occurs in the training process of producing similar or identical samples, which fails to capture the full diversity of the data distribution. It typically happens when the model converges to a limited set of data patterns, ignoring the variety of the actual data, which will decrease the generalization ability of the model and affect its practical application effect. The results generated by VAE may lack detailed information, often leading to blurred results [28].

In contrast, recent advancements in diffusion models (DMs) have demonstrated that employing principled probabilistic diffusion modeling can yield high-quality mapping from randomly sampled Gaussian noise to complex target distribution, without suffering from mode collapse or instabilities. The foundations of the diffusion model can be traced back to the pioneering work in 2015 [29], which was inspired by nonequilibrium thermodynamics. This concept has been further developed and popularized in subsequent works, such as denoising diffusion probabilistic models (DDPM) [30], improved denoising diffusion probabilistic models (IDDPM) [31] and denoising diffusion implicit models (DDIM) [32].

Taking DDPM as an example, DMs typically encompass two processes: the forward diffusion process and the reverse diffusion process, as shown in Figure 2, both of which are characterized by a T-step Markov chain.

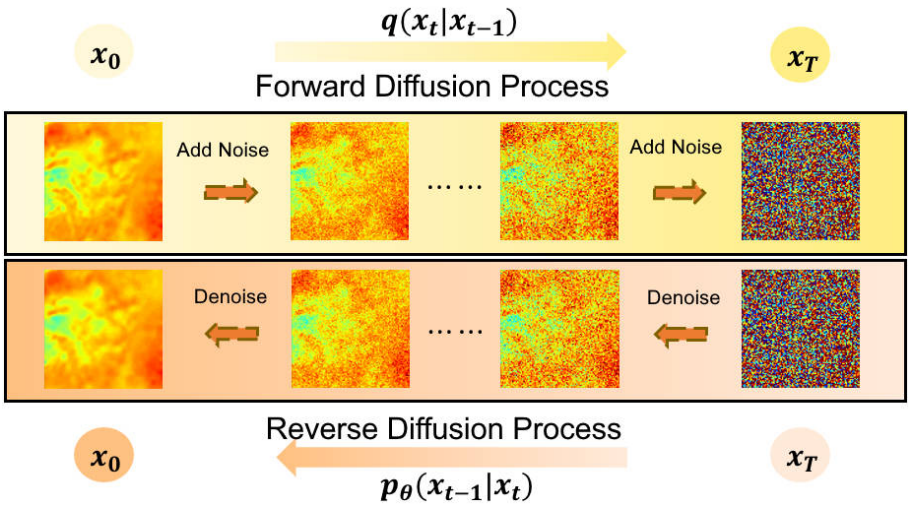

Figure 2. The forward and reverse diffusion processes of diffusion model, where $q(x_t|x_{t-1})$ means the forward process that transforms distribution $q(x_{t-1})$ to $q(x_t)$ and $p_\theta(x_{t-1}|x_t)$ represents the reverse process that transforms distribution $p_\theta(x_t)$ to $p_\theta(x_{t-1})$.

The forward diffusion process transforms the start data distribution into a final Gaussian distribution. Firstly, the initial data x_0 are defined, and then Gaussian noise is progressively added in each timestep until it reaches pure Gaussian noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at timestep T by T iterations. During each mid-timestep t, noisy data x_t are generated with the same shape as x_0. The noise incorporated in the diffusion process is specified by a predefined sequence of $\beta_{1:T} \in (0,1]^T$. Denote $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{n=1}^{T} \alpha_n$, and $\beta_1 < \beta_2 < \ldots < \beta_t (t \in [1, T])$, each iteration of the forward diffusion process can be described as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

where the process that transforms distribution $q(x_0)$ to $q(x_t)$ can be condensed into one single step:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha_t}}x_0, (1 - \overline{\alpha_t})\mathbf{I}) \tag{3}$$

Thus, x_t can be directly sampled as follows:

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{(1 - \overline{\alpha_t})}\xi, \xi \in \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

Meanwhile, the reverse process focuses on learning the inverse of the forward diffusion process and generating a distribution that resembles real data distribution. The reverse diffusion process first samples a random noise $x_T \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then gradually denoises it until it reaches a high-quality output x_0. Define $p_\theta(x_t)$ as the data distribution at timestep t in the reverse process, and a neural network $\epsilon_\theta(x_t, t)$ is involved in predicting the uncertain variables, where θ represents the network parameters. Each iteration of the reverse diffusion process can be described as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}) \tag{5}$$

where $\mu_\theta(x_t, t)$ is the mean value computed using Equation (6):

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \epsilon_\theta\frac{1 - \alpha_t}{\sqrt{1 - \overline{\alpha_t}}}) \tag{6}$$

and σ_t^2 is the variance value calculated using Equation (7):

$$\sigma_t^2 = \frac{1 - \overline{\alpha_{t-1}}}{1 - \overline{\alpha_t}}\beta_t \tag{7}$$

In addition, there is only one unknown variable that should be learned in the reverse process. DMs use a neural network $\epsilon_\theta(x_t, t)$ to estimate it. To train the model of $\epsilon_\theta(x_t, t)$, a timestep t and a noise $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ are randomly sampled to generate noisy data x_t with the given real data x_0, according to Equation (2). Then, the entire network parameters are optimized by the following loss function:

$$Loss = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(x_t, t)\right\|_2^2\right] \tag{8}$$

Substituting x_t, i.e., Equation (4) into Equation (8), yields the following:

$$Loss = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha_t}}x_0 + \sqrt{(1 - \overline{\alpha_t})}\xi, t)\right\|_2^2\right] \tag{9}$$

In the DMs designed for single-image SR, some models directly generate high-resolution images in the image domain [20], carrying out the forward and reverse diffusion process on the input pixel space. These models demand excessive iteration steps (about 100–1000 steps) on large-scale denoising models to precisely capture data details, which consumes massive computational resources [23]. Furthermore, directly conducting the diffusion process in a high-dimensional data space requires a large amount of GPU memory and training times.

Other models conduct diffusion processes in the latent space. DMs are utilized to generate low-dimensional latent variables. These latent variables are then employed for SR reconstruction, like [33] and DIFFIR. These models offer efficiency advantages by circumventing diffusion processes in high-dimensional data spaces. For the sake of training resources and efficiency, we selected DIFFIR as the base architecture model.

2.3. DIFFDS Method

Computer vision SR tasks focus mainly on the diversity of visual perception. However, when using DMs for SST spatial downscaling, the focus is more on accurately generating meso- or small-scale dynamic SST processes than on their diversity. This is crucial because the downscaled high-resolution SST data may later be used for other downstream tasks like forecasting. In such cases, inaccurate processes can negatively impact the validity of task results.

However, applying the original DIFFIR to SST spatial downscaling can lead to inauthentic details in the downscaled results, resulting in poor subjective and objective evaluations. To ensure that the downscaled results are more consistent with the destination high-resolution SST data and to enhance the accuracy of the reconstructed data, the DIFFDS method is proposed.

DIFFDS consists of three stages: the first stage, training a dynamic IRformer (DIRformer) and a compact prior extraction network (CPEN); the second stage, training the denoising network and finally, the third stage, also the inference stage, where the trained networks are used for the downscaling task. The workflow (illustrated in Figure 3) of DIFFDS employs DDPM to generate a guidance vector called compact IR prior representation (IPR) and then utilizes IPR to direct the DIRformer in the downscaling process. During the training phase, CPEN and DIRformer are first trained. The trained CPEN is integrated into the forward diffusion process of DDPM to train the denoising network. Once the denoising network is trained, the IPR can be predicted through the reverse sampling process of DDPM. Subsequently, guided by the predicted IPR, the pre-trained DIRformer performs the downscaling process.

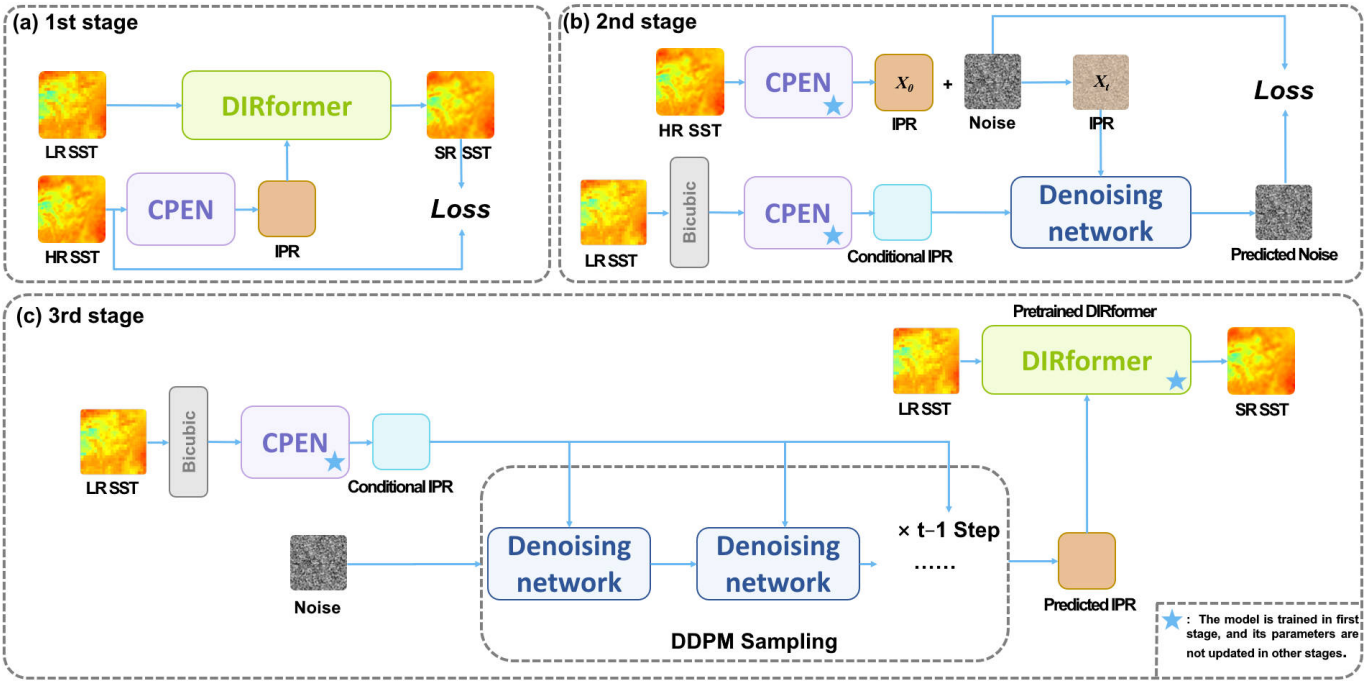

Figure 3. (**a**) illustrates the first training stage, detailing the training processes of CPEN and DIRformer. (**b**) depicts the second training stage, which is also the forward process of DDPM. The 3rd stage (**c**) shows the inference process of DIFFDS.

In DIFFDS, we redesign the structure of the transformer block in DIRformer by incorporating cross-attention and channel-attention mechanisms. Channel attention focuses on increasing the weight of certain significant feature channels [34] while suppressing others within the data. This selective enhancement allows DIFFDS to maintain some SST texture anomalies. Additionally, cross-attention enables DIFFDS to learn the underlying distribution structures in low-resolution SSTs. This ensures that the produced high-resolution SST closely aligns with the original data, minimizing deviations. These adjustments have markedly improved the performance of the model. Specifically, DIFFDS has primarily modified the transformer block, while keeping the rest of the network framework consistent with DIFFIR. Figure 4 depicts the entire architecture.

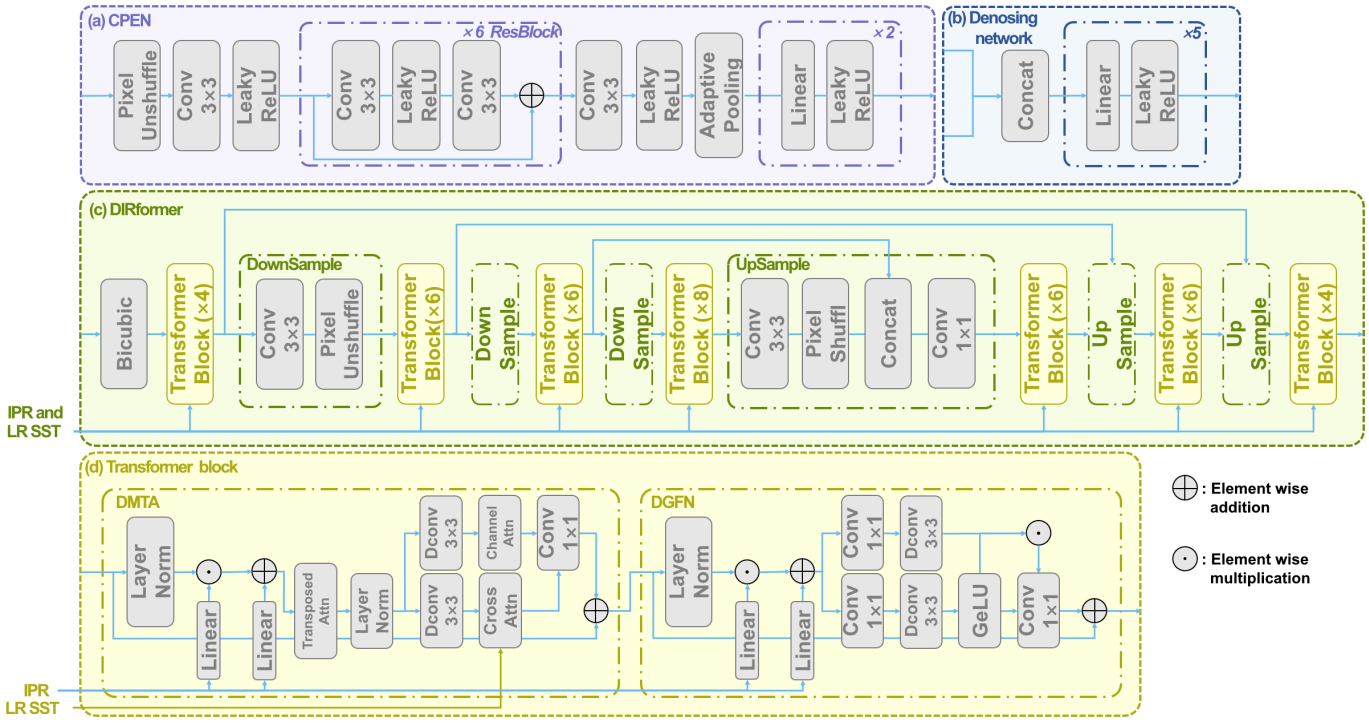

Figure 4. Architecture details of DIFFDS. (**a**) CPEN, (**b**) Denoising network, (**c**) DIRformer, (**d**) Transformer block.

2.3.1. CPEN

With the input of high-resolution SSTs, CPEN learns to generate a low-dimensional IPR, which can guide the DIRformer in the SST downscaling process.

As described in Figure 4a, the CPEN is constructed by a series of residual blocks, convolution layers, and linear layers. First, the original high-resolution SST training data are processed by a pixel unshuffle layer to facilitate training speed. Then, a 3×3 convolution layer and a Leaky ReLU are applied to extract the feature map. After that, multiple residual blocks are utilized to calculate and refine feature representations from the current feature map. Finally, IPR is produced through the transformation of an adaptive pooling layer and several linear and Leaky ReLU layers. The CPEN process can be described as follows.

$$IPR = CPEN(SST_{HR}) \tag{10}$$

2.3.2. DIRformer

Under the guidance of IPR, DIRformer accepts the input low-resolution SST and then generates the corresponding high-resolution versions (as shown in Figure 3a). The DIRformer is constructed by stacking transformer blocks in a U-Net structure. Each of these transformer blocks comprises a modified dynamic multi-head transposed attention (DMTA) part and a dynamic gated feed-forward network (DGFN) module (Figure 4d).

After the CPEN process, IPR information is obtained, denoted as X_0. As provided for dynamic modulation parameters, X_0 is designed into the DMTA of DIRformer to integrate the current feature map with the high-resolution SST feature information and guide the downscaling process of the DIRformer. As shown in Equation (11), F is the input feature map of DMTA:

$$F' = W_l^1 X_0 \odot Norm(F) + W_l^2 X_0 \tag{11}$$

where $\odot$ indicates element-wise multiplication, *Norm* denotes layer normalization, W_l represents linear layer, and F' are the output feature map, respectively.

Next, the original DIRformer employs a transposed multi-head attention mechanism (a kind of efficient muti-head attention [35]) to process F', so the DMTA process of DIFFIR is as follows:

$$\hat{F} = TransposedAttn(F') + F \tag{12}$$

To improve its performance, a 3×3 depth-wise convolution layer is introduced to capture more intricate and detailed spatial features. The extracted feature map is then separately fed into a channel-attention layer and a cross-attention layer. In practice, the channel attention layer is utilized to enhance the representational power of neural networks by emphasizing informative features in vital channels while suppressing incorrect or noisy information channels, allowing DMTA to facilitate more effective utilization of the guidance information generated by the DDPM. Meanwhile, the cross-attention layer is added to fuse low-resolution SST context information and the current feature map. This enhancement can make the results more consistent with the original SST structure, and avoid generating abnormal dynamical processes. Finally, the results are integrated using a 1×1 convolution layer. The DMTA process of DIFFDS can be described as follows:

$$F'' = TransposedAttn(F') \tag{13}$$

$$\hat{F} = W_c(CrossAttn(W_d^1(F'')) + ChannelAttn(W_d^2(F''))) + F \tag{14}$$

where W_c and W_d are the convolution layer and the depth-wise convolution layer.

In the DGFN process, the same feature map F' can be obtained by Equation (11), which integrates IPR information with the DGFN's input feature map F. Then, a 1×1 convolution unit is exploited to aggregate information from different channels. Next, a 3×3 depth-wise convolution unit is added to aggregate information from spatially neighboring pixels. Besides, the gating mechanism is adopted to enhance information encoding. The overall process of DGFN is defined in Equation (15), $\hat{F}$ is the output of DGFN.

$$\hat{F} = GELU(W_d^1 W_c^1 F') \odot W_d^2 W_c^2 F' + F \tag{15}$$

2.3.3. Denoising Network

The structure of the denoising network consists of multiple stacked linear layers and Leaky ReLU layers (Figure 4b). It concatenates the noisy IPR and the reference conditional IPR in the channel dimension as inputs and then produces the previous timestep noise as output.

2.3.4. Training and Inference

The training of DIFFDS incorporates two phases: the first phase trains CPEN and DIRformer, and the second phase is the forward diffusion process of DDPM, which trains the denoising network.

In the first stage, as listed in Algorithm 1, CPEN and DIRformer are trained together, which can make CPEN learn to transform high-resolution SST data into IPR and force DIRformer to reconstruct high-resolution SST guided by IPR. For each pair of low- and high-resolution SST data, the high-resolution SST is sent to CPEN to obtain IPR, after which IPR and low-resolution SST are sent to the DIRformer to generate downscaled SST. The training loss function in phase 1 is defined as:

$$Loss1 = \left\| SST_{HR} - SST_{SR} \right\|_1 + L_{adv}(SST_{HR}, SST_{SR}) \tag{16}$$

where SST_{HR} and SST_{SR} are the ground-truth and downscaled high-resolution data, respectively. $\| \cdot \|_1$ denotes the L1 norm, and L_{adv} is the adversarial loss used in Real-ESRGAN [36].

Algorithm 1 Training CPEN and DIRformer

Input: low-resolution SST SST_{LR} and high-resolution SST SST_{HR}
1: **for** SST_{LR}, SST_{HR} **do**
2: $X_0 = CPEN(SST_{HR})$
3: $SST_{SR} = DIRformer(X_0, SST_{LR})$
4: Calculate and optimize *Loss*1
5: **end for** if *Loss*1 converges
Output: Trained CPEN and DIRformer

Phase 2 (as shown in Algorithm 2) is the forward diffusion process of DDPM. In this stage, the denoising network is trained to predict noise. After preparing parameters like $\overline{\alpha_t}$ and α_t, for each pair of low- and high-resolution SST data, the pre-trained CPEN extracts IPR X_0 from high-resolution SSTs. Then, X_t is sampled by the forward diffusion equation, according to Equation (4). After substituting variables X_0 and X_t into Equation (4), the formula is obtained as follows:

$$X_t = \sqrt{\overline{\alpha_t}} X_0 + \sqrt{(1 - \overline{\alpha_t})}\xi, \xi \in \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{17}$$

where the $\overline{\alpha_t}$ is the same parameter in Equation (4), and ξ is a Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Algorithm 2 Training DDPM

Input: Trained CPEN, $\beta_{1:T} \in (0, 1]^T$, low-resolution SST SST_{LR} and high-resolution SST SST_{HR}
1: Init: $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{n=1}^{t} \alpha_n$
2: **for** SST_{LR}, SST_{HR} **do**
3: $X_0 = CPEN(SST_{HR})$
4: Sample a $t \in [1, T]$
5: Sample X_t
6: $X_c = CPEN(Bicubic(SST_{LR}))$
7: Calculate and optimize *Loss*2
8: **end for** if *Loss*2 converges
Output: Trained Denoising network

Using bicubic interpolation, the SST_{LR} data are upsampled to the same resolution as SST_{HR}, after which it is sent to CPEN to earn condition IPR X_c. This IPR X_c functions as conditional information, facilitating the denoising network's capability to accurately forecast noise patterns. Finally, the parameters of the denoising network can be updated according to the DDPM loss function. In Loss Equation (18), ϵ_θ is the denoising network, ϵ is the Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and t is the timestep, *Concat* is the concatenation step:

$$Loss2 = \left\| \epsilon - \epsilon_\theta(Concat(X_c, X_t), t) \right\|_2^2 \tag{18}$$

For the inference process as illustrated in Algorithm 3, a random noise sample is initialized. Eventually, the reverse diffusion process Equation (19) is deprived of the predicted IPR X_0, which contains the corresponding high-resolution SST information.

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}}(X_t - \epsilon_\theta(Concat(X_c, X_t), t)\frac{1 - \alpha_t}{\sqrt{1 - \overline{\alpha_t}}}) + \xi\sqrt{\frac{1 - \overline{\alpha_{t-1}}}{1 - \overline{\alpha_t}}\beta_t} \tag{19}$$

where β_t, α_t, $\overline{\alpha_t}$ are the same parameters in Equations (6) and (7), t is the sampled timestep, and ϵ_θ is the denoising network.

Subsequently, X_0 and the low-resolution SST input are passed to the pre-trained DIRformer for spatial downscaling, and then the DIRformer outputs the spatially downscaled high-resolution SST data.

Algorithm 3 Inference

Input: Trained CPEN, DIRformer, $\epsilon_\theta, \beta_{1:T} \in (0,1]^T$, low-resolution SST SST_{LR}
1: Init: $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{n=1}^{t} \alpha_n$
2: **for** SST_{LR} **do**
3: $X_c = CPEN(Bicubic(SST_{LR}))$
4: Sample $X_t \in \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: **for** $t = T, \ldots, 1$ **do**
6: Sample $\xi \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$ else $\xi = 0$
7: Update X_{t-1} by Equation (19)
8: **end for**
9: $SST_{SR} = DIRformer(X_0, SST_{LR})$
10: **end for**
Output: Downscaled results SST_{SR}

2.4. Evaluation Metrics

To thoroughly evaluate the quality and accuracy of the proposed DIFFDS method as well as other baseline models, we have selected a comprehensive set of objective evaluation metrics, including root mean square error (RMSE), mean absolute error (MAE), Bias, peak signal-to-noise ratio (PSNR) and temporal correlation coefficient (TCC). These metrics will measure the differences between spatial resolution improved SST values and actual high-resolution SST values.

In this paper, after obtaining the downscaling results, the data are first denormalized from the range $(-1, 1)$ back to the normal range. RMSE, Bias, MAE, and TCC are calculated directly using these denormalized data. To calculate the PSNR, the data are first normalized to (0, 1) by dividing by 35.

3. Experiments and Results

3.1. Experiments Design

Based on the dataset in Section 2.1.2, we conduct 4× scale downscaling experiments. For comparison, the selected baseline models include Lasso regression, Bicubic, RCAN, ESRGAN and DIFFIR. All models were trained on the aforementioned OSTIA dataset. Since the SST high-resolution results obtained by Bicubic interpolation do not completely align with the edges of the ground truth, we calculate the metrics related to Bicubic using only the overlapping portion of the Bicubic results and the ground truth.

Regarding DIFFDS, in training stage 1, we set the number of transformer blocks per layer in DIRformer to [4, 6, 6, 8], and the number of attention heads per layer to [1, 2, 4, 8]. The number of resblocks in CPEN is set to 6. In training stage 2, the reverse sampling steps of DDPM are set to 200, with a beta scheduler configured to linear, and beta start and beta end values (the β_1 and β_T in Section 2.2) set to 0.0001 and 0.02, respectively. The timestep spacing algorithm is set to leading [37]. The learning rate for both stage 1 and stage 2 starts at 5×10^{-5}, with a cosine scheduler for the learning rate, and the batch size is 16 for both phases.

For the training of DIFFDS, we utilized an NVIDIA RTX 4070Ti graphics (Lenovo, Beijing, China) card equipped with 12 GB of memory. Under these specific hardware conditions, the training process necessitated a total of 400 epochs. Each epoch required approximately 4 min to complete. Upon the completion of the training phase, the model's inference time for performing downscaling operations was recorded to be around 1.8 s. While this setup was sufficient to accomplish training objectives, we observed that training speeds could be further optimized. Based on our practical experience, we recommend

employing a graphics card with at least 16 GB of memory. This would likely result in more efficient training processes and a reduced overall training time.

3.2. Results

3.2.1. Metrics Evaluation

As listed in Table 1, the proposed DIFFDS method outperforms other models across various objective evaluation metrics. It achieves an average RMSE of 0.1074 °C, an average Bias of −0.0043 °C, an average MAE of 0.0654 °C, an average PSNR of 50.48 dB and a TCC of 0.9610, demonstrating high precision and effectiveness.

With respect to RMSE, DIFFDS excels in all aspects, including average, maximum, and minimum values. RCAN, ESRGAN, and DIFFIR followed with a 0.02 °C gap. Bicubic performs worse than the above models. The Lasso regression method shows poor performance with a mean of 0.3347 °C, which shows a noticeable gap compared to other models. This highlights the effectiveness of deep learning models in addressing downscaling problems.

For MAE, DIFFDS again takes the lead in average, maximum, and minimum values, followed by RCAN. The differences among the remaining deep learning models are relatively insignificant. Lasso regression and Bicubic show a significant disparity compared to deep learning models on this metric. Regarding Bias, DIFFIR achieves the best average value of −0.0003 °C, ESRGAN performs the best in maximum values at 0.0008 °C, and RCAN performs the best in minimum values at −0.0023 °C, respectively.

As for PSNR, DIFFDS consistently delivers the best results in average, maximum, and minimum values, demonstrating that the downscaling results have very low noise. It is followed by ESRGAN, RCAN, and DIFFIR, while Lasso regression and Bicubic lag behind.

In terms of TCC, DIFFIR achieves the highest value of 0.9634, indicating that the downscaled results exhibit good temporal consistency compared to the true values. DIFFDS, ESRGAN, RCAN and Lasso have lower TCC values than DIFFIR, at 0.9610, 0.9536, 0.9444 and 0.9615, respectively. Bicubic displays the lowest TCC, around 0.88. Although DIFFDS shows a slightly reduced TCC compared to DIFFIR, it still outperforms other benchmark models. An obvious fact is that Lasso regression's TCC is comparable to most deep learning models, indicating that Lasso has lower accuracy regarding reconstruction error and reconstruction quality, but has stronger temporal correlation and can capture the temporal sequence characteristics of SST. Deep learning models, on the other hand, have higher accuracy in reconstructing SSTs in the spatial dimension but have weaker temporal correlation, which may be because deep learning models focus more on capturing spatial features and, to some extent, neglect the information in the temporal dimension.

In general, the improved DIFFDS presents further performance improvements in most metrics, highlighting its superiority over DIFFIR and other baseline models.

Figure 5a shows that DIFFDS surpasses other neural networks in terms of the maximum value, min value, median and upper and lower quartiles. This is consistent with the results presented in Table 1.

The evaluation in Figure 5b,c is largely similar to that in Figure 5a, showing the consistency of DIFFDS across different metrics.

For the Bias box plot in Figure 5d, DIFFIR and DIFFDS outperform other models in terms of the median, and the upper and lower quartiles. ESRGAN performs best in maximum values, while RCAN excels in minimum values. Regarding the data distribution between the maximum and minimum values, and the interquartile, Lasso, DIFFDS, and DIFFIR take the lead. RCAN and ESRAGN have relatively poor performance, whereas Bicubic displays the largest fluctuations. As for DIFFDS, with its modified network architecture, although it is slightly inferior to DIFFIR, it still outperforms most of the comparison models. In the TCC plot, DIFFIR surpasses the other models in maximum, minimum, and median values. DIFFDS and Lasso follow closely, showing only minor differences from DIFFIR. This indicates these models' results have good temporal consistency with the ground truth and are capable of accurately capturing SST change trends over time.

ESRGAN has relatively poor TCC distributions, while RCAN and Bicubic perform the worst in terms of minimum and median values.

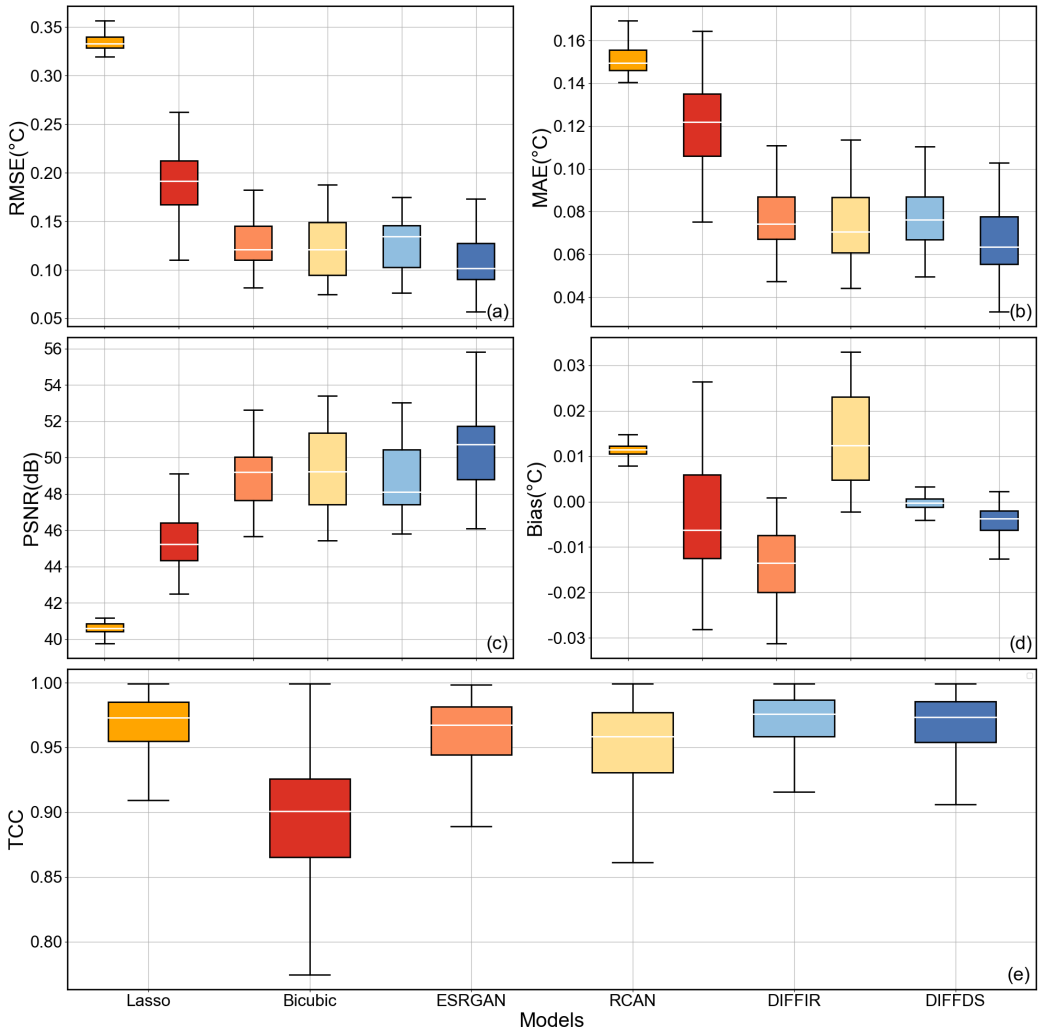

Figure 5. The maximum, minimum, median, both upper and lower quartiles of each metric, (**a**) RMSE, (**b**) MAE, (**c**) PSNR, (**d**) Bias, for each method. (**e**) is the TCC plot for each point in the experiment sea area.

Table 1. The average/maximum/minimum value of RMSE, MAE, Bias, PSNR, and the value of TCC.

Model	RMSE (°C)	MAE (°C)	Bias (°C)	PSNR (dB)	TCC
Lasso	0.3347/0.3644/0.3196	0.1506/0.1692/0.1405	0.0113/0.0155/0.0056	40.60/41.19/39.69	0.9615
Bicubic	0.1891/0.1750/0.0762	0.1206/0.1645/0.0753	−0.0039/0.0264/−0.0282	45.44/50.06/42.50	0.8858
ESRGAN	0.1259/0.1824/0.0819	0.0763/0.1109/0.0474	−0.0138/0.0008/−0.0312	48.99/52.61/45.67	0.9536
RCAN	0.1224/0.1875/0.0747	0.0735/0.1136/0.0442	0.0139/0.0329/−0.0023	49.39/53.41/45.42	0.9444
DIFFIR	0.1269/0.1750/0.0761	0.0770/0.1105/0.0495	−0.0003/0.0040/−0.0066	48.77/53.04/45.82	0.9634
DIFFDS	0.1074/0.1734/0.0567	0.0654/0.1027/0.0331	−0.0043/0.0023/−0.0145	50.48/55.87/46.10	0.9610

3.2.2. Analysis of Temporal Trends

The temporal variations in the metrics reveal similar general trends across deep learning models. In particular, neural network models tend to perform relatively poorly during spring and summer, while their performance improves in autumn and winter. This seasonal pattern may be attributed to the influence of the East Asian monsoon in the study

area. During the spring and summer, the ocean–atmosphere interaction intensifies and is more active than in autumn and winter. Factors such as heat flux, evaporation, and advection processes on the sea surface can cause rapid changes in SST. Additionally, warm seasons are often accompanied by stronger solar short-wave radiation and convective activities, resulting in larger SST fluctuation. These fluctuations pose a challenge for model representations, contributing to the relatively poor performance of neural network models during this period.

Unlike spring and summer, SST changes in autumn and winter are relatively stable and exhibit simpler textures. This makes it easier for deep learning models to capture accurate SST representations and features, resulting in better performance during these seasons.

The Lasso regression method exhibits a relatively stable trend without showing significant seasonal variations, compared with other methods. This stability may be attributed to its regularization property, which reduces model complexity and results in more consistent performance across different seasons.

The results for DIFFIR are not ideal due to the lack of specialized adjustments for SST spatial downscaling tasks (Figure 6a). Specifically, DIFFIR performs worse during the summer period compared to other models, as SST fluctuations are more significant during this time, making it more challenging for DIFFIR to learn. In such cases, the unmodified network architecture is more prone to generate some erroneous content, resulting in larger RMSE values. For ESRGAN, its RMSE is slightly smaller than that of DIFFIR. This is because ESRGAN has a relatively simple generator architecture, and incorporates adversarial and perceptual losses. These factors compel GAN to balance different losses during the training process. Consequently, the generated content may contain more detailed textures, but the RMSE remains relatively high. RCAN, which utilizes a channel-attention mechanism, performs better than ESRGAN and DIFFIR in terms of RMSE. However, its relatively simple network structure limits its performance.

The performance of DIFFDS stands out, as it achieves the lowest average, maximum, and minimum RMSE values. The overall RMSE curve reveals that DIFFDS consistently achieves the lowest RMSE on most dates. However, during the latter part of autumn and winter (November 2021–March 2022), RCAN tends to perform better. This suggests that during periods of relatively stable SST variations and simple SST distribution structures, the downscaling performance of DIFFDS is not as pronounced. Other models can also achieve similar results.

DIFFDS generally exhibits lower error than DIFFIR, suggesting that the improved network structure effectively corrects the downscaling results. For ESRGAN, its RMSE gap during the autumn and winter seasons is relatively higher than that of other models, differing from the performance of RCAN. This discrepancy may be attributed to its inherent characteristics, making it less sensitive to data variations in these seasons.

The results for MAE (Figure 6b,c) are similar to RMSE: DIFFDS achieves the best results on most dates. However, in autumn and winter, the differences between DIFFDS and other models tend to be small or even reverse, suggesting that the performance gap narrows during these seasons. The Lasso regression method consistently performs the worst on these metrics.

In terms of Bias variations (Figure 6d), DIFFIR shows the most stable trend, closely followed by DIFFDS and Lasso. This indicates that the predictions of these methods are statistically close to the true values, without systematic overestimation or underestimation of SST over long-term averages. The errors are balanced in both positive and negative directions, resulting in a very small overall error trend. In contrast, RCAN, ESRGAN and Bicubic exhibit greater fluctuations in Bias variation, and their numerical values for Bias are comparatively poorer.

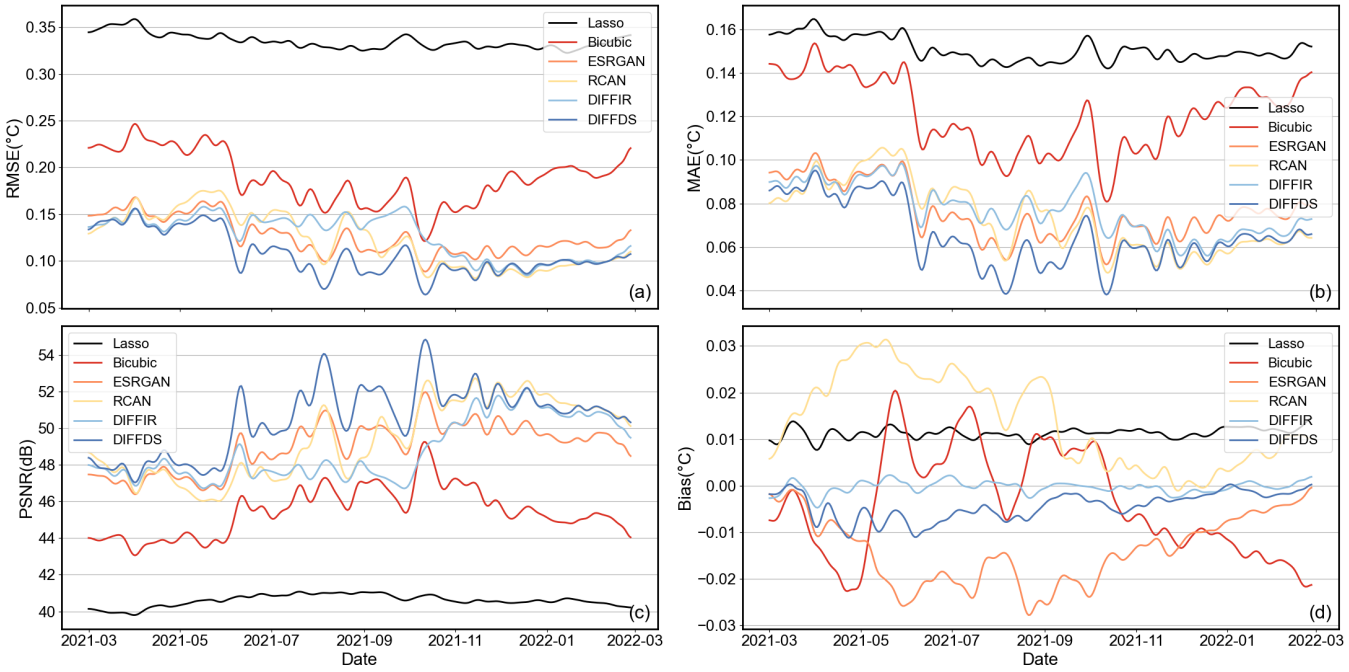

Figure 6. The time series variations for each metric, (**a**) RMSE, (**b**) MAE, (**c**) PSNR, (**d**) Bias, from March 2021 to February 2022. The text highlighted in blue marks the date of the dotted line. The results of these two dates will be used in the discussion section.

3.2.3. Correlation Analysis

Analysis of the correlation distribution across all data points revealed that the results from these neural network models closely approximate the true values (Figure 7). The fitting curves of baseline models almost coincide with the 1:1 line, except for Lasso, which exhibits a relatively noticeable deviation. This suggests that the vast majority of data point predictions from the aforementioned models are accurate, and the deep learning spatial downscaling methods can effectively correct the deviation between low- and high-resolution SST points. Notably, DIFFDS achieves a correlation coefficient of 0.9981, which is the closest to 1, reflecting the highest degree of consistency between its results and the ground truth SST values. The other baselines follow closely, with correlation coefficients very similar to that of DIFFDS, indicating their excellent performance as well.

For Bicubic and Lasso regression methods, the distribution of data points (especially between 22 °C and 32 °C) shows a more pronounced deviation compared to other methods. This is because these methods do not adequately consider the spatial distribution relationship of each data point with its surrounding data points during the downscaling process, resulting in larger errors. For other deep learning models, the distribution of data points between 28 °C and 30.5 °C shows a noticeable deviation, while those above 30.5 °C align well. This could be because SST data points above 30.5 °C make up a smaller proportion, specifically 4 percent of the total data points according to our computation. Additionally, SST data points above 30.5 °C tend to occur in more homogeneous regions (e.g., consistently warm waters), which are easier for the neural network to learn. In contrast, data points in the 28–30.5 °C range constitute a larger portion, accounting for 64 percent of the total data points. These SST points are often significantly impacted by monsoons, exhibiting distinct fluctuations during different periods, making reconstruction quite challenging. Consequently, the number of points where the downscaling results do not match the true values is relatively higher, leading to greater discrepancies in the scatter plot.

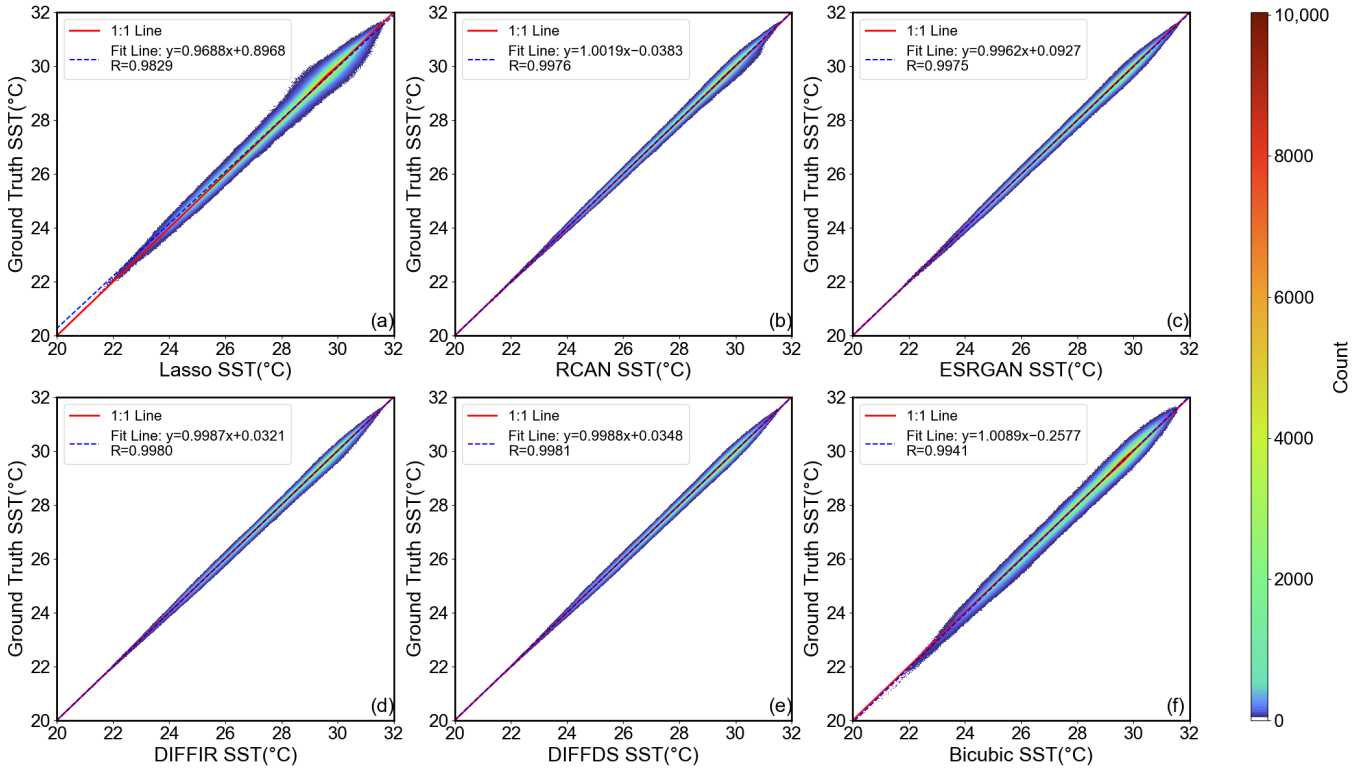

Figure 7. (**a**–**f**) The density scatter plots of each model.

Furthermore, we randomly selected a 100-day sample subset from the test set to calculate the correlation coefficient, repeating this process 100 times. As shown in the following Table 2, the correlations are similar to those discussed above. Although the correlation coefficient differences between our DIFFDS method and other algorithms are relatively small, it obviously demonstrates that these methods can reserve the large-scale structure from the LR inputs. The reconstruction quality of small-scale structures can evaluate the effectiveness of different algorithms in terms of RMSE, MAE, and PSNR metrics, as shown in Table 1.

Table 2. The correlation coefficient test results.

Model	Mean Correlation	Standard Deviation
Lasso	0.9821	0.0269
Bicubic	0.9942	0.0019
ESRGAN	0.9973	0.0012
RCAN	0.9974	0.0010
DIFFIR	0.9978	0.0011
DIFFDS	0.9981	0.0008

4. Discussion

4.1. Specific Samples Examination

In this section, several samples are selected to analyze the downscaling results and to compare the differences among those models. The first sample is on 29 June 2021, belonging to the summer season (Figure 8). The SST distribution during this period is relatively complex, with many dynamic processes. As shown in the previous time series variations in metrics, summer samples tend to accentuate the differences between models.

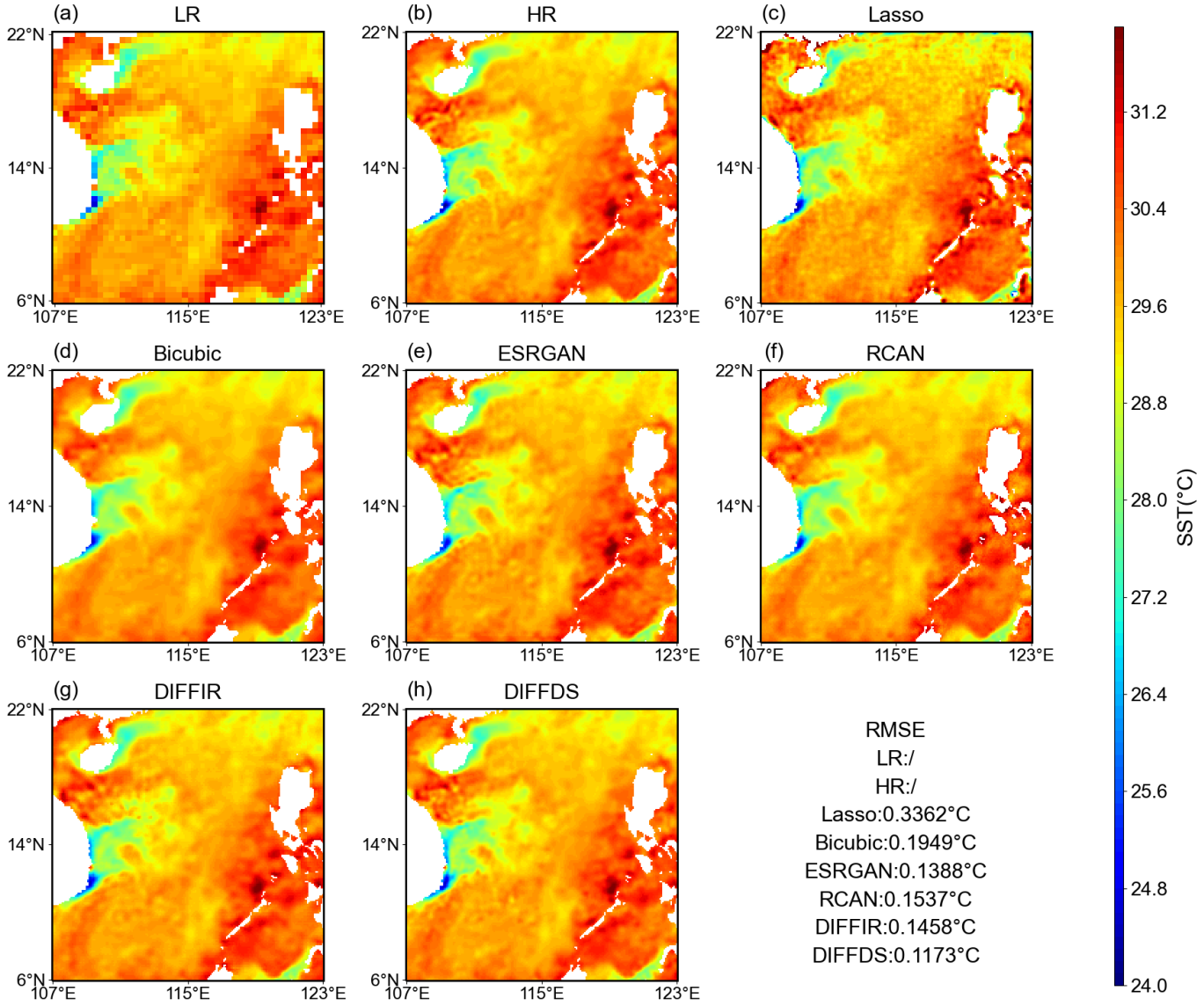

Figure 8. The SST distribution on 29 June 2021 of each model, these eight subplots individually display the high-resolution SST, low-resolution SST, and the downscaled results of each model along with their RMSE (°C).

Firstly, in terms of RMSE, DIFFDS and ESRGAN emerge as the top two models, with the others trailing behind. From Figure 8, we can find that all methods can reconstruct the basic patterns of high-resolution SSTs. However, upon closer inspection, RCAN exhibits relatively blurry and smooth SST structures, similar to Bicubic, whereas ESRGAN, DIFFIR, and DIFFDS reveal more complex reconstructed SST structures. This disparity can be attributed to the fact that models using single L1 loss as a loss function may perform well in metrics such as RMSE and PSNR. However, L1 loss tends to erase high-frequency information [38] during training, making the final downscaled high-resolution SST appear like the smoothed version of ground truth.

In contrast, ESRGAN uses additional adversarial and perceptual losses during training, which enables the generation of richer information while suppressing RMSE and other metrics. The results of the Lasso regression also exhibit relatively complex patterns, but they are not consistent with the true values because the Lasso method does not adequately consider the spatial distribution relationship of each data point with its surrounding data points during the downscaling process, resulting in a higher RMSE.

Generative diffusion models such as DIFFIR, and DIFFDS utilize their distribution fitting capabilities to produce realistic high-resolution SST samples. These models are easier to train than GANs. As a result, they perform better across various metrics in the final downscaled results, showing their potential to generate high-quality SST samples.

Based on the Bias map (Figure 9), the relatively low errors observed in all deep learning models across most marine regions underscore their ability to tackle spatial downscaling problems. However, an acute bias area emerges near the land, specifically between 107°E–114°E and 10°N–17°N, where the models' performance is compromised. As seen in Figure 8, the SST variation in this region is large, indicating a higher level of downscaling complexity compared to other areas. This leads to large deviations in the downscaled results. For Lasso regression, because it did not reconstruct the high-resolution SST features well, its bias regions are relatively larger compared to those of the deep learning model. Compared to Bicubic, deep learning models demonstrate greater accuracy across most areas due to their complex structure and powerful learning capabilities, resulting in lower bias.

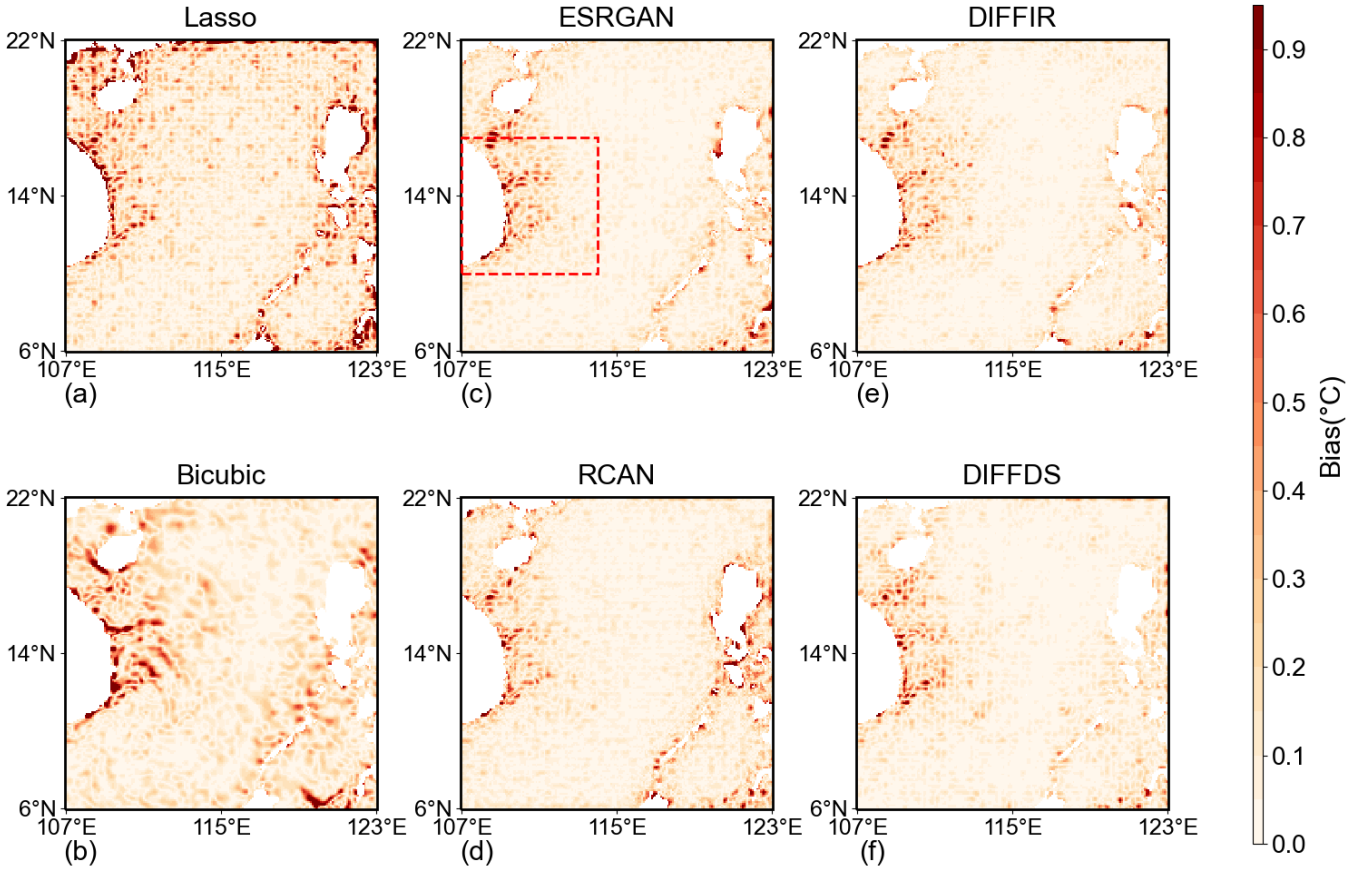

Figure 9. (**a**–**f**) The absolute Bias map between ground truth and each deep learning model on 29 June 2021. The red box displays the intense Bias area.

As shown in Figure 10, the region from 10°N to 18°N and 107°E to 115°E is selected to examine at a basin scale. Although some discrepancies exist, the DIFFDS-generated results more accurately capture most of the dynamic processes, yielding superior downscaled results. This improvement can be attributed to the adjusted network architecture, which leverages the guidance information provided by the diffusion model while also considering the overall SST distribution structure inherent in low-resolution SSTs.

The original DIFFIR without cross-attention primarily focuses on the guidance information from DDPM, somewhat neglecting the overall SST structure. This oversight results in poor SST structures. For the DIFFIR results (Figure 10), erroneous SST content is shown in the red-boxed area of Figure 10g. These textures cannot be found in the red-boxed area of

the input low-resolution SSTs (Figure 10a). Conversely, the corresponding area in DIFFDS is corrected with no obvious anomalous textures. This enhanced structural effectiveness can be directly attributed to the introduction of channel-attention and cross-attention. For ESRGAN, although it generates rich details in downscaled high-resolution SSTs, it still fails to recover certain features, such as the high-temperature area in the block box of subplot Figure 10e.

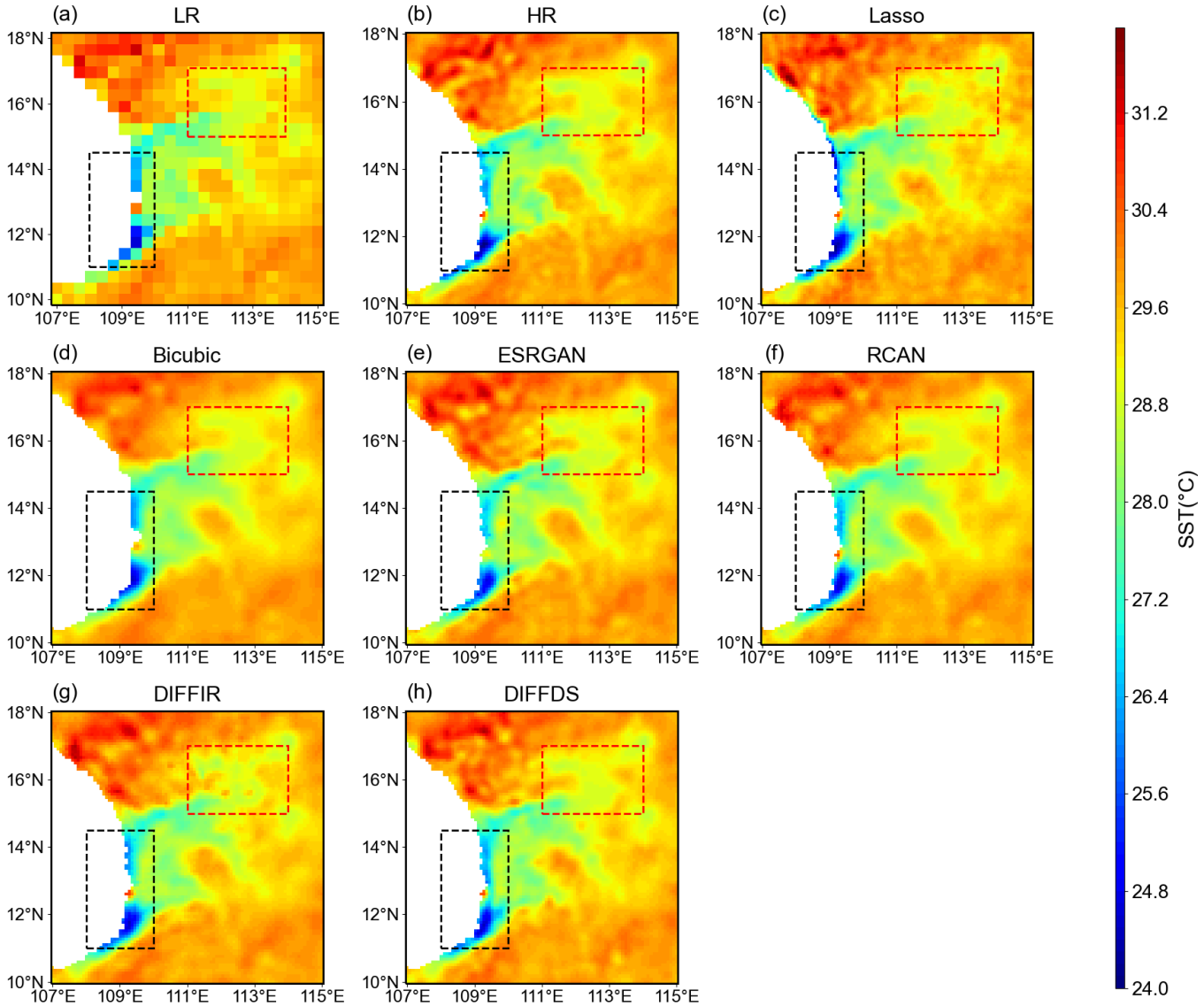

Figure 10. The results on 29 June 2021 zoomed from 9°N–17°N and 108°E–116°E. The red box and black box areas display erroneous SST contents.

Next, we analyze another sample on 1 August 2021. The overall SST pattern on this day is relatively simpler, so the reconstruction task for all deep learning models is less challenging. However, the downscaled results of the Lasso regression show a considerable discrepancy compared to the deep learning methods, with an RMSE of 0.3284 °C, which is higher than that of neural networks. As illustrated in Figure 11, the performance differences among neural networks are further reduced. Their downscaling results are quite similar, effectively restoring SST patterns in this region. This similarity in performance can also be found in their Bias maps (Figure 12). Except for Lasso regression, the bias distribution of these models in this sea area is relatively uniform, with no highly concentrated error regions. This indicates that deep learning models can recover most of the mesoscale dynamic

processes in the SST downscaling task, but they still have some difficulty capturing certain small-scale processes.

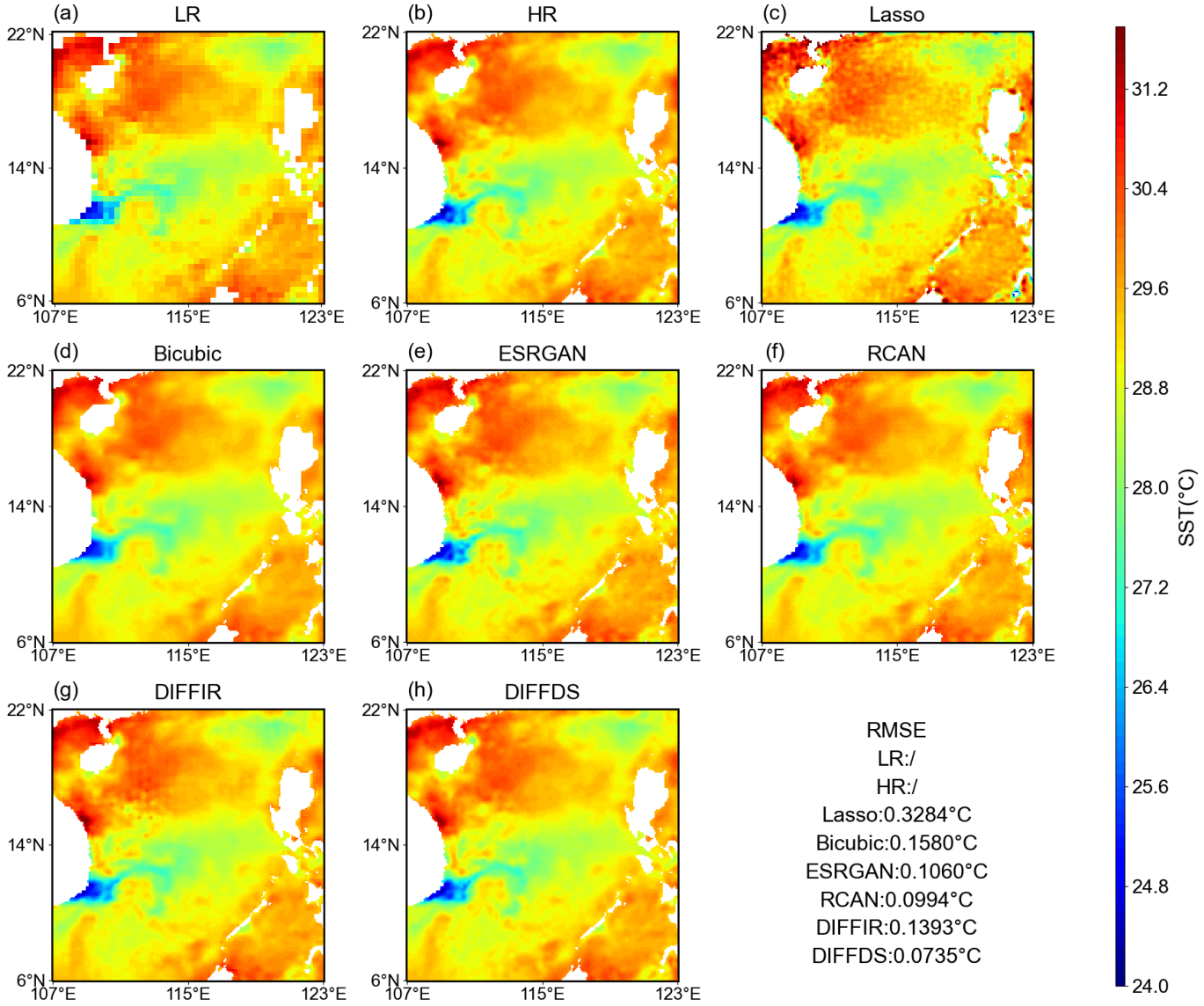

Figure 11. The SST distribution on 1 August 2021 of each model, these eight subplots individually display the high-resolution SSTs, low-resolution SSTs, and the downscaled results of each model along with their RMSE (°C).

The absence of previously significant bias in the 107°E–114°E area, 10°N–17°N area (red box in Figure 9) is likely due to the relatively stable fluctuations in SST in that region. Regarding RMSE, all deep learning models show a noticeable decrease compared to the sample on 29 June 2021, indicating further improvements in the plain SST structure environments. In this context, DIFFDS achieves the lowest RMSE at 0.0735 °C, reflecting the precision of the reconstructed high-resolution SSTs compared to other baselines. The RMSEs of DIFFIR and RCAN are close to those of DIFFDS. As shown in Figure 13, the results from deep learning models show visually minor differences (in the black box area) but are closer to the true SST distribution compared to the Bicubic and Lasso regression methods.

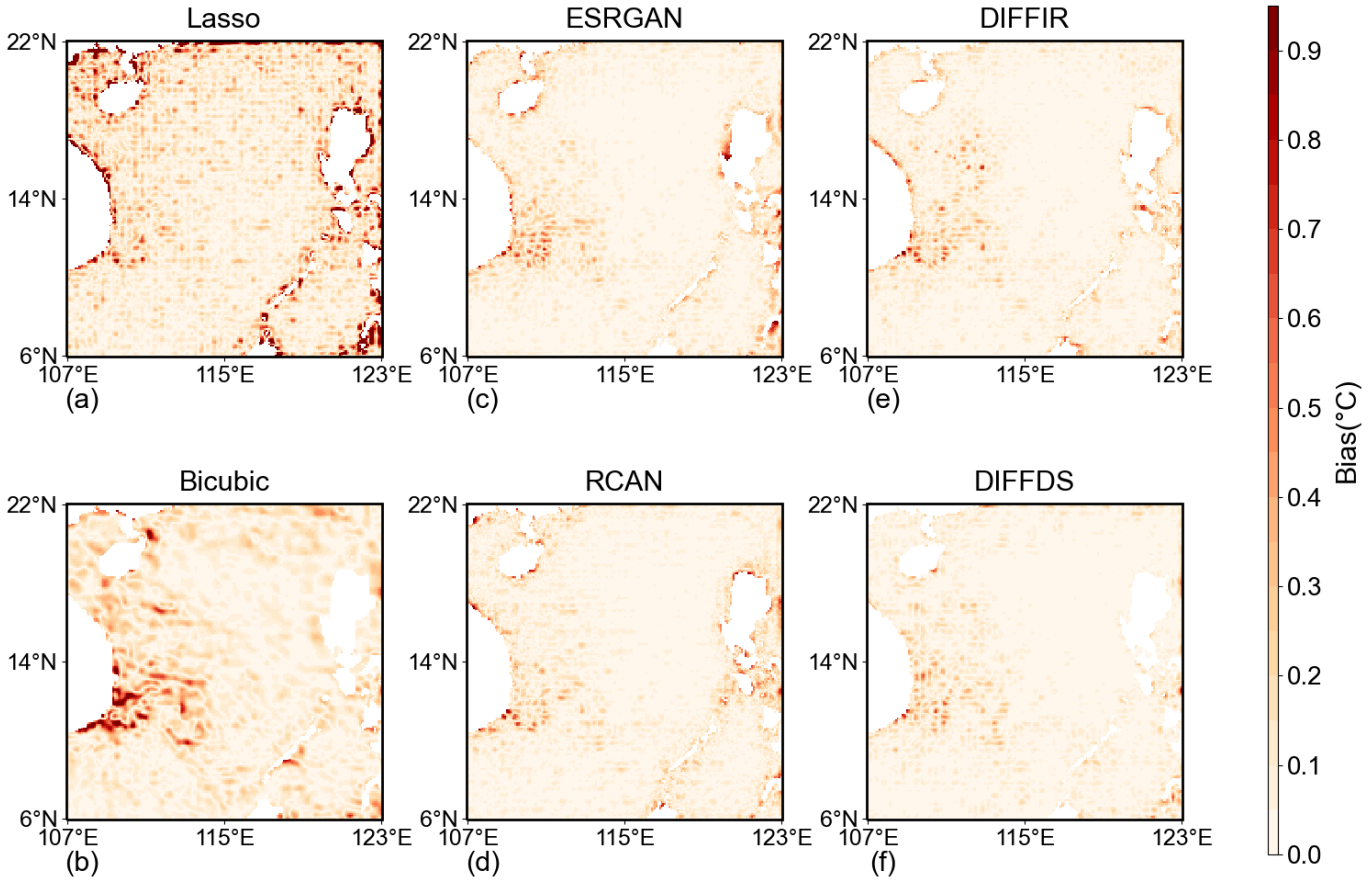

Figure 12. (**a**–**f**) The absolute Bias map between ground truth and each deep learning model on 1 August 2021.

4.2. Further Comparison of DIFFDS and DIFFIR

In the preceding sections, we analyzed the improvements in DIFFDS over the original DIFFIR using various metrics and specific downscaling results. In this section, to further illustrate the performance differences between these two models, we selected a case study from 27 June to 1 July 2021, in regions where coastal upwelling in Vietnam can be found.

In Figure 14, the coastal upwelling phenomenon in the HR exhibits a continuous variation process, which can be observed from the shape of the red 27 °C isotherms. However, this is not evident in the low-resolution SST data. As seen in the figures, the coastal upwelling morphology (the shape of red isotherms) in the low-resolution SST data over these five days shows almost no differences and fails to reflect a continuous morphological change. For DIFFIR, the downscaled results can not accurately represent this variability across those days. Its upwelling results are similar to those from Bicubic, being merely a simple magnification of the low-resolution data. If there are no significant changes in the upwelling morphology in the LR data, then the results also show no significant changes.

Based on the results of DIFFIR on 27 June, the upwelling morphology is relatively close to the true SST values. However, its upwelling morphology showed little change in the following days, remaining essentially the same as the previous day. It cannot exhibit a noticeable variation, as observed in the ground truth. This contributes to the suboptimal performance of DIFFIR's downscaling results.

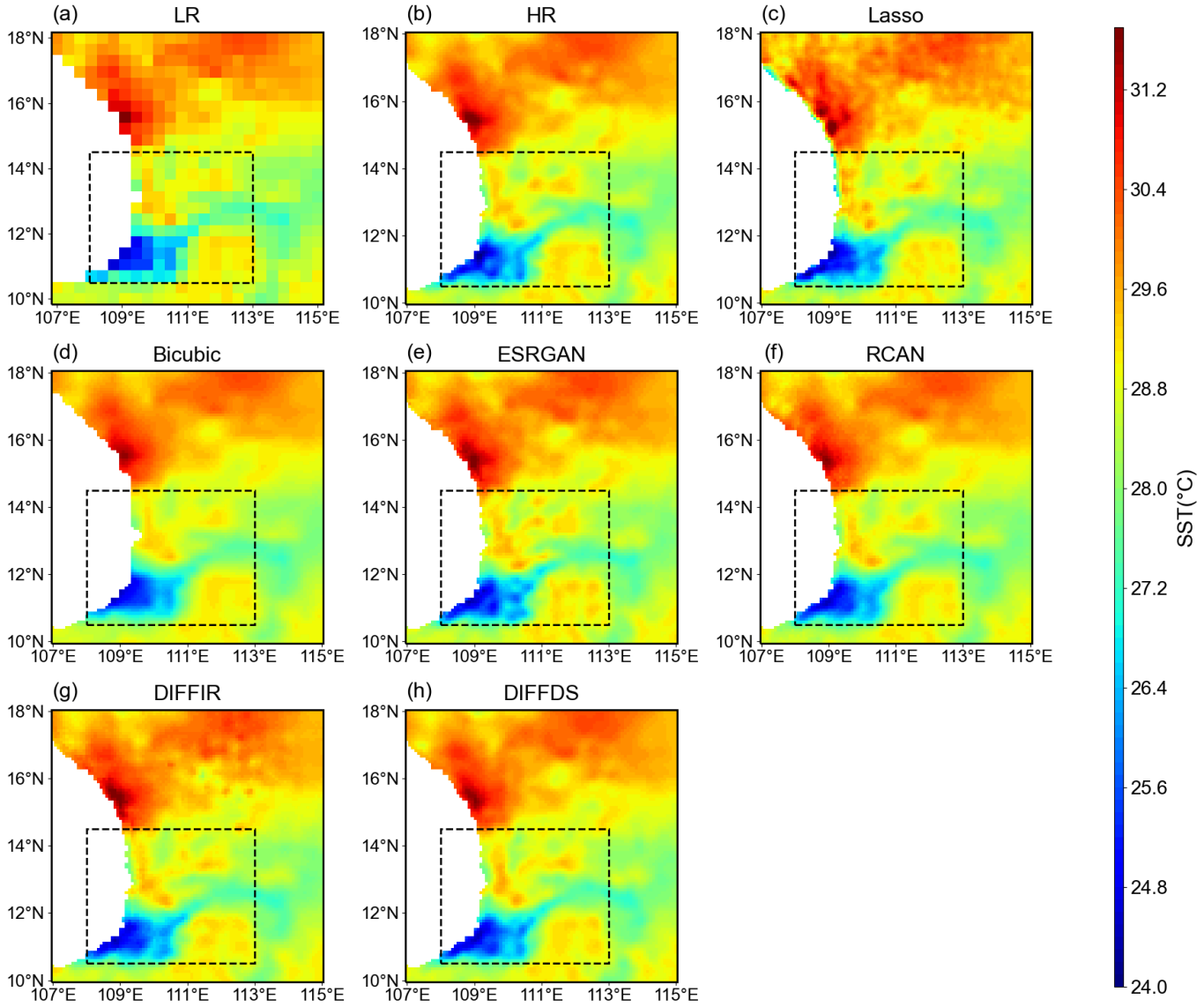

Figure 13. The results on 1 August 2021 zoomed from 10°N–18°N to 107°E–115°E. The black box area displays complex SST contents.

In contrast, the downscaled coastal upwelling from the DIFFDS displays a more pronounced variation over these five days, closely mirroring the ground truth. Additionally, the daily upwelling patterns produced by the DIFFDS are more aligned with the ground truth. This indicates that an improved network structure is more effective at learning the inherent SST distribution patterns in low-resolution SSTs by incorporating channel attention and cross-attention. It captures and reflects the subtle variations within the low-resolution SST data, and then translates them into high-resolution SSTs accurately. Consequently, the final downscaled results are more consistent with the ground truth, making the DIFFDS more suitable for SST spatial downscaling tasks.

In Figure 15, we present an analysis of the spatial distribution RMSE for both DIFFIR and DIFFDS on the test set. The figure clearly illustrates that DIFFDS consistently exhibits a lower RMSE across a majority of the examined regions when compared to DIFFIR. This trend is particularly pronounced within the area highlighted by the red box. In this specific region, DIFFIR demonstrates a significantly higher RMSE, indicating a poorer performance. However, the corresponding region in DIFFDS shows a markedly improved RMSE, highlighting the correction and enhancement achieved by our method.

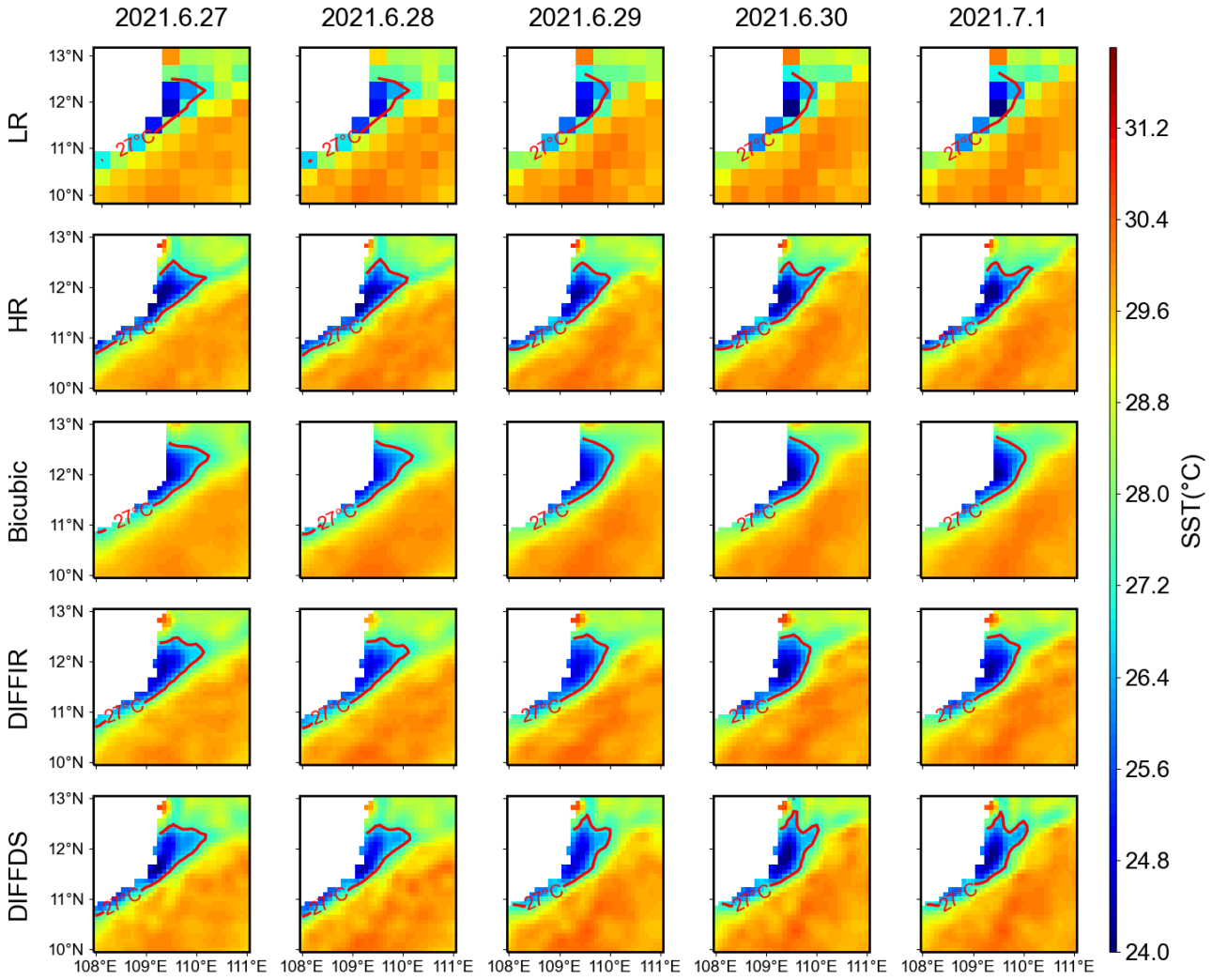

Figure 14. Variations in the low, ground truth, DIFFIR, and DIFFDS SST data over a continuous five-day period. Red isotherms are used to highlight the boundary of the upwelling currents.

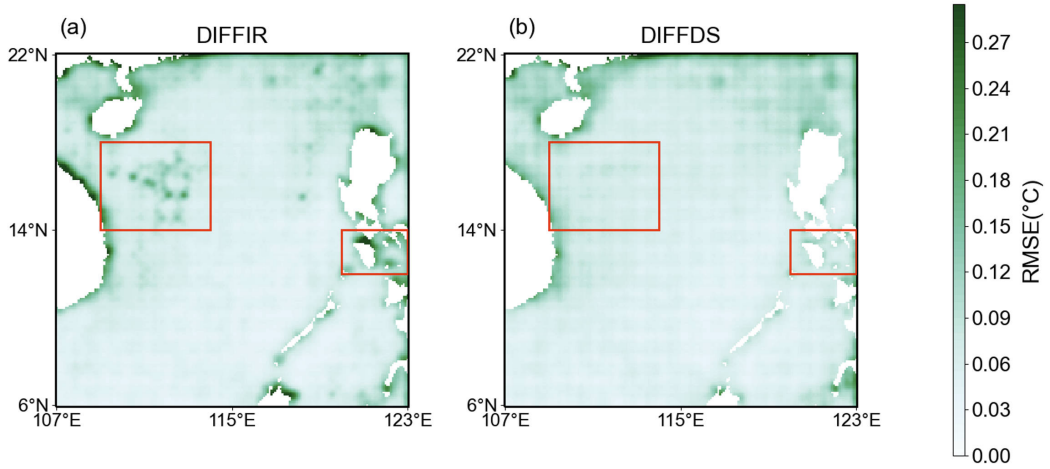

Figure 15. The spatial distribution of RMSE for DIFFIR and DIFFDS on the test set. The red boxed area highlights the significant difference between DIFFIR and DIFFDS.

The improvements observed underline the advantages of the enhanced DIFFDS method over the original DIFFIR approach. The reduction in RMSE across various spatial

zones suggests that DIFFDS not only improves overall accuracy but also corrects specific areas of high error, making it a more reliable and robust solution.

4.3. Challenges with High Variability

In Section 3.2.2, we found that the intense fluctuations in SSTs during spring and summer pose a significant challenge to the model's downscaling capabilities. This can ultimately affect the accuracy of the downscaled results. Despite utilizing nearly three decades of data for training, DIFFDS has yet to effectively capture underlying SST patterns under conditions of high variability. We attribute this limitation to several factors. Firstly, recent years have witnessed abnormal changes in weather systems that may not be fully represented in the historical data used for training. Past experiences may not always apply to current or future situations, thereby restricting the model's performance in such cases. Additionally, the inherent limitations of our model's architecture and the complexity of the ocean system may also contribute to this shortcoming. Moreover, relying solely on SSTs as an input variable has its limitations.

When the region is more homogeneous, it is economically feasible to obtain the downscaling reconstruction using simple interpolation (usually Bicubic interpolation). However, the real-world system is very complicated, for instance, extreme ocean events and large changes in ocean elements (high variability situations). Therefore, there is an urgent need for efficient and effective downscaling algorithms that can respond extremely well to real-world systems, which is also the goal of our future work. All of the downscaling methods, including Bicubic interpolation, can greatly preserve the large-scale structure from the low-resolution inputs in terms of correlation coefficient differences. Additionally, the visual results and the corresponding zoomed-in regions, as well as several quantitative metrics (as shown in the figures and tables above), mainly evaluate the reconstruction quality of small-scale processes among different algorithms. On dates with high variability, or during periods when SST variability is relatively stable, our DIFFDS model surpasses the baseline models in most of these metrics, demonstrating that DIFFDS can effectively reconstruct more high-resolution activities under various conditions.

To improve downscaling results during periods of high variability, incorporating other oceanic factors, such as salinity, and atmospheric factors, like surface wind speed, into the neural network is essential. This approach could enable the model to capture a more comprehensive representation of the underlying dynamics and enhance its downscaling capabilities.

5. Conclusions

This study proposes a novel spatial downscaling model (DIFFDS) and applies it to SST spatial downscaling to explore its potential possibility and relieve the growing demand for high-resolution oceanic data. The results of our experiment demonstrate its effectiveness for spatial downscaling of SSTs. The proposed DIFFDS can restore some meso-scale processes that disappeared in low-resolution SSTs, generating accurate results.

To enhance the consistency and accuracy of the downscaled results, we designed a modified DMTA layer by introducing channel-attention and cross-attention mechanisms. The redesign enables the model to extract precise and effective information from the guidance IPR provided by the diffusion model, while also considering the overall SST distribution structure inherent in low-resolution SST data. This approach allows the model to suppress abnormal SST textures in the downscaled results, resulting in more realistic and accurate outputs.

Furthermore, we compared the performance of DIFFDS with commonly used CNN, GAN, and regression methods, to highlight its superiority over other models. Experimental results indicate that DIFFDS achieves an average RMSE of 0.1074 °C and PSNR of 50.48 dB in the 4× scale downscaling task. Meanwhile, the generated content is comparable to the raw high-resolution SST data, such as the coastal upwelling along Vietnam.

Future work will focus on refining this method by incorporating other oceanic or atmospheric elements or integrating physical mechanisms into the model. This will further

improve the interpretability and effectiveness of the DIFFDS, enabling more accurate and reliable SST spatial downscaling.

Author Contributions: Methodology, S.W.; experiments, S.W.; writing–original draft, S.W.; data curation, X.Z.; project administration, X.Z.; writing–review, X.Z.; mathematical analysis, X.L.; writing–review and editing, X.L.; resources, S.G. and J.L.; software, S.G. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (No. SML2023SP202 and No. SML2023SP219), and the Zhuhai Basic and Applied Basic Research Foundation (No. 2320004002806).

Data Availability Statement: The OSTIA dataset in this paper is available to download in Copernicus at the following addresses: https://data.marine.copernicus.eu/product/SST_GLO_SST_L4_REP_OBSERVATIONS_010_011/description, accessed on 1 August 2024.

Acknowledgments: We acknowledge Natural Earth @naturalearthdata.com for providing the region image in Figure 1.

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

References

1. Pastor, F. Sea Surface Temperature: From Observation to Applications. *J. Mar. Sci. Eng.* **2021**, *9*, 1284. [CrossRef]
2. Huang, X.; Rhoades, A.M.; Ullrich, P.A.; Zarzycki, C.M. An evaluation of the variable-resolution CESM for modeling California's climate. *J. Adv. Model. Earth Syst.* **2016**, *8*, 345–369. [CrossRef]
3. Shen, Z.; Shi, C.; Shen, R.; Tie, R.; Ge, L. Spatial Downscaling of Near-Surface Air Temperature Based on Deep Learning Cross-Attention Mechanism. *Remote Sens.* **2023**, *15*, 5084. [CrossRef]
4. Perez, J.; Menendez, M.; Camus, P.; Mendez, F.J.; Losada, I.J. Statistical multi-model climate projections of surface ocean waves in Europe. *Ocean Model.* **2015**, *96*, 161–170. [CrossRef]
5. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef] [PubMed]
6. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
7. Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
8. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image Super-Resolution Using Dense Skip Connections. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
9. Dong, X.; Xi, Z.; Sun, X.; Gao, L. Transferred Multi-Perception Attention Networks for Remote Sensing Image Super-Resolution. *Remote Sens.* **2019**, *11*, 2857. [CrossRef]
10. Salvetti, F.; Mazzia, V.; Khaliq, A.; Chiaberge, M. Multi-Image Super Resolution of Remotely Sensed Images Using Residual Attention Deep Neural Networks. *Remote Sens.* **2020**, *12*, 2207. [CrossRef]
11. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
12. Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; Zeng, T. Transformer for Single Image Super-Resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 18–24 June 2022; pp. 456–465. [CrossRef]
13. Conde, M.V.; Choi, U.J.; Burchi, M.; Timofte, R.; Swin2SR: SwinV2 Transformer for Compressed Image Super-Resolution and Restoration. In *Computer Vision—ECCV 2022 Workshops*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 669–687. [CrossRef]
14. Ducournau, A.; Fablet, R. Deep learning for ocean remote sensing: An application of convolutional neural networks for super-resolution on satellite-derived SST data. In Proceedings of the 2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS), Cancun, Mexico, 4 December 2016. [CrossRef]
15. Khoo, J.J.D.; Lim, K.H.; Pang, P.K. Deep Learning Super Resolution of Sea Surface Temperature on South China Sea. In Proceedings of the 2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Miri, Sarawak, Malaysia, 26–28 October 2022. [CrossRef]
16. Izumi, T.; Amagasaki, M.; Ishida, K.; Kiyama, M. Super-resolution of sea surface temperature with convolutional neural network- and generative adversarial network-based methods. *J. Water Clim. Chang.* **2022**, *13*, 1673–1683. [CrossRef]
17. Zou, R.; Wei, L.; Guan, L. Super Resolution of Satellite-Derived Sea Surface Temperature Using a Transformer-Based Model. *Remote Sens.* **2023**, *15*, 5376. [CrossRef]

18. Saharia, C.; Chan, W.; Saxena, S.; Lit, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Gontijo-Lopes, R.; et al. Photorealistic text-to-image diffusion models with deep language understanding. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; pp. 36479–36494.
19. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, arXiv:2204.06125.
20. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image Super-Resolution Via Iterative Refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4713–4726. [CrossRef] [PubMed]
21. Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; Chen, Y. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* **2022**, *479*, 47–59. [CrossRef]
22. Shang, S.; Shan, Z.; Liu, G.; Wang, L.; Wang, X.; Zhang, Z.; Zhang, J. Resdiff: Combining cnn and diffusion model for image super-resolution. In Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024.
23. Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Van Gool, L. DiffIR: Efficient Diffusion Model for Image Restoration. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
24. Stark, J.D.; Donlon, C.J.; Martin, M.J.; McCulloch, M.E. OSTIA: An operational, high resolution, real time, global sea surface temperature analysis system. In Proceedings of the OCEANS 2007-Europe, Aberdeen, Scotland, 18–21 June 2007. [CrossRef]
25. Donlon, C.J.; Martin, M.; Stark, J.; Roberts-Jones, J.; Fiedler, E.; Wimmer, W. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.* **2012**, *116*, 140–158. [CrossRef]
26. Good, S.; Fiedler, E.; Mao, C.; Martin, M.J.; Maycock, A.; Reid, R.; Roberts-Jones, J.; Searle, T.; Waters, J.; While, J.; et al. The Current Configuration of the OSTIA System for Operational Production of Foundation Sea Surface Temperature and Ice Concentration Analyses. *Remote Sens.* **2020**, *12*, 720. [CrossRef]
27. Liu, K.; Qiu, G.; Tang, W.; Zhou, F. Spectral Regularization for Combating Mode Collapse in GANs. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
28. Huang, H.; Li, Z.; He, R.; Sun, Z.; Tan, T. IntroVAE: Introspective variational autoencoders for photographic image synthesis. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018.
29. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 2256–2265.
30. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020; pp. 6840–6851.
31. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8162–8171.
32. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
33. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
35. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
36. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montréal, QC, Canada, 10–17 October 2021; pp. 1905–1914.
37. Lin, S.; Liu, B.; Li, J.; Yang, X. Common Diffusion Noise Schedules and Sample Steps are Flawed. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024. [CrossRef]
38. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.