Hyper-parameters Effects in Conditional Diffusion Models for Accurate Sea Surface Temperature Reconstruction

Muhammad Sarmad, Emanuele Mele, Rajat Srivastava, Marco Pulimeno, Massimo Cafaro, and Italo Epicoco

Department of Engineering for Innovation, University of Salento
Lecce-73047, Italy

Abstract. Accurate representation of oceanic conditions is fundamental for reliable climate modeling, weather forecasting, and environmental monitoring. However, ocean models and observational datasets often exhibit systematic biases due to limitations in model physics, parameterizations, resolution, or observational coverage. In this work, we propose a diffusion model for bias correction. We systematically evaluated its performance for Sea Surface Temperature on the oceanic sea surface temperature generation by varying different hyperparameters in the U-Net architecture. The model is trained to denoise simulated data and reconstruct the SST field guided by reanalysis data. Our results demonstrate that increasing the base channel's depth significantly improves the model's performance, with improvements in convergence speed, reconstruction accuracy, and spatial detail retention. Quantitative metrics such as root mean squared error (RMSE), Pearson's correlation coefficient (PCC), and coefficient of determination (R^2) show notable gains up to a base channel depth of 64, beyond which performance gains plateau. A detailed temporal generalization analysis using seasonal batches every two months confirms the robustness of the model in varying SST regimes. At the same time, qualitative visualizations show sharp and coherent reconstructions with minimal error. The study highlights the trade-off between model complexity and performance and identifies 64 base channels as a computationally efficient and accurate configuration for SST modeling using diffusion-based generative methods.

Keywords: Diffusion Models, Oceanic Dataset, Architectural Parameters, Bias Correction

1 Introduction

Generative modeling has become a significant and advancing area of research in recent years. Key model types, including generative adversarial networks (GANs) [1], variational autoencoders (VAEs) [2], auto-regressive models [3], flow models [4] [5], and diffusion models [6][7], have made great improvements. These models have been successfully applied to a variety of tasks, such as generating realistic images [8][9][10], image super-resolution [11][12][13], image editing [14] [15], and text-to-image generation [16] [17][18].
Generative models have revolutionized its applications in computer vision by enabling the creation of realistic and diverse images. The evolution of generative models has progressed from early probabilistic frameworks to the powerful diffusion models used today. Initially, Variational Autoencoders (VAEs) [2] provided probabilistic foundations but often generated blurry outputs. Generative Adversarial Networks (GANs) [1] followed, setting a new benchmark for visual quality, but instability and mode collapse hindered their reliability. To address those challenges, diffusion models have emerged as an alternative class of generative models inspired by non-equilibrium thermodynamics. These models gradually transform data into noise through a forward diffusion process and then reconstruct it through a reverse denoising process. This iterative approach enables diffusion models to generate diverse, high-fidelity samples while maintaining stable training dynamics, effectively overcoming the key limitations of VAEs and GANs. The development

David C. Wyld et al. (Eds): Signal, CDKP, AdNLP, MLT, SAI – 2025
pp. 79-98, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150906

of diffusion-based generative models began with early work on score-based generative modeling and stochastic differential equations. Sohl-Dickstein et al. [6] proposed the concept of a diffusion process for generative modeling, where data is progressively corrupted through the addition of noise, and this corruption can be reversed using a learned denoising process. However, this approach remained computationally intensive until Ho et al. [7] proposed the Denoising Diffusion Probabilistic Model (DDPM) architecture. DDPMs employ a two-stage process, where noise is progressively added to data (forward process) and then subsequently removed (reverse process). Their architecture demonstrated that it is possible to train a straightforward U-Net architecture to predict noise addition and generate high-quality images, rivaling state-of-the-art GANs. This method allows for highly stable training and impressive image generation quality, showing that slow, iterative sampling can outperform the adversarial dynamics of GANs in terms of fidelity.

Subsequent research aimed at the efficiency of DDPMs. Song et al. [19] in 2020 presented Score-Based Generative Modeling using stochastic differential equations (SDEs), uniting score-based and diffusion models into a unified probabilistic framework. This was an enhancement of the flexibility of diffusion models and connected them with a broader class of generative models. Although DDPMs were efficient, their sampling was still computationally demanding. In 2021, Denoising Diffusion Implicit Models (DDIMs) addressed this bottleneck by accelerating inference using a non-Markovian approach, reducing the sampling steps while maintaining most of DDPM's output quality, rendering diffusion models more realistic [20]. Other innovations above DDPMs in 2021, including an improved noise schedule and enhanced model architecture [18], resulted in monumental quality gains, further pushing diffusion models to state-of-the-art. At the same time, in 2022, Latent Diffusion Models (LDMs) [21] shifted the diffusion process into a compressed latent space, reducing memory and computational cost while preserving image quality, thus making scalable high-resolution synthesis possible. Conditional generation has advanced further with Ho and Salimans [22] introducing classifier-free guidance and allowing DDPMs to generate controllable outputs without the necessity for extra classifiers. This approach has been widely adopted in text-to-image models such as DALLE-2 [17], Imagen [16], and Stable Diffusion [21]. Despite these successes, a critical gap remains in understanding how architectural and training parameters influence performance in diffusion models. The key hyper-parameters, which include the number of base channels in the U-Net, the learning rate, and the number of diffusion steps, play a significant role in determining both the quality and efficiency of generated outputs. However, their impacts have not been investigated in a comprehensive and organized manner, particularly in scientific applications where data will always be limited and high fidelity matters.

In this study, we present a comprehensive analysis of architectural and training parameters in diffusion models tailored to correct the bias of oceanic simulated data, obtained by numerical models, reconstructing data near the reanalysis and hence near the observations. For the parameters study, we focused only on the Sea Surface Temperature (SST) fields from oceanographic datasets, but the bias correction approach using the diffusion model can also be used for other variables such as salinity, sea surface height, and velocity. Capturing fine-scale spatial features with high accuracy, especially in coastal areas, is essential for effective ocean forecasting and the management of marine resources. These regions are characterized by sharp temperature gradients and are heavily influenced by localized currents, freshwater inputs, and complex topography. Inaccuracies in representing these features can lead to substantial errors in downstream applications like habitat modeling, predicting coastal upwelling, or detecting marine heatwaves. Enhancing reconstruction ac-

curacy in these zones, therefore, improves the reliability of high-resolution ocean models used for operational forecasting and informed decision-making.
In contrast to prior work, which often uses fixed architectural choices borrowed from natural image generation, our study conducts a focused analysis on the base channel size, training steps, and learning rate parameters. This makes it one of the few efforts to adapt the diffusion model design systematically for geophysical data assimilation and reconstruction tasks. Through extensive experiments, we identify trade-offs between fidelity, stability, and computational cost and provide practical insights into optimal configurations. We make three key contributions: (i) we present a systematic evaluation of architectural parameters, in particular the base channel width, and analyze their impact on the accuracy of the reconstruction of SST fields; (ii) we conduct an empirical analysis of learning rate and diffusion step variations, shedding light on their influence on training speed, model stability and generalization capabilities; (iii) we propose practical guidelines for selecting diffusion model configurations that effectively balance the model performance with computational efficiency, offering valuable direction for the application of generative models for bias correction. By addressing the interplay between architecture and training parameters, this work aims to guide the effective design of conditional diffusion models for bias correction in the weather and climate domain.

The structure of the paper is outlined as follows. Section 2 presents a comprehensive background on denoising diffusion probabilistic models, with particular emphasis on the conditional formulation adopted in this study. Section 3 describes the dataset, the architecture of the proposed conditional diffusion model, the training methodology, and the evaluation metrics. Section 4 discusses the experimental results, including quantitative assessments, visual analyses, and ablation studies exploring key parameters such as base channel size, learning rate, and training duration. Finally, Section 5 summarizes the main findings, highlights key insights, and outlines potential directions for future research.

2 Background

2.1 Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPM) [7] are a class of generative models that learn to generate images by sequentially removing noise from an initially perturbed input image. The training process involves a forward diffusion stage, where Gaussian noise is incrementally added to an image over multiple time steps, transforming it into pure noise. The generative process then learns to reverse this degradation by denoising the sample step by step, generating a clean image. This reverse process is modeled as a Markovian diffusion process, where the model gradually refines the noisy input, starting from white noise and progressively generating a coherent image. Diffusion models are based on two complementary processes, the forward diffusion process and the reverse denoising process. Together, these processes enable the generation of complex data distributions, such as images, from simple noise distributions.

Conditional Diffusion Model Conditional diffusion models extend the framework of denoising diffusion probabilistic models (DDPMs) by incorporating external information to guide the data generation process. In contrast to unconditional diffusion models that sample purely from noise, conditional models generate outputs that are both near to the target clean images and consistent with an auxiliary input. In this study, we employ a conditional diffusion model to reconstruct high-resolution sea surface temperature (SST)

fields from model simulations, using reanalysis SST data as the reference target. The core of the model lies in learning to reverse a fixed forward process, where Gaussian noise is gradually added to the ground truth SST field x_0 over a series of timesteps $t = 1, 2, \ldots, T$. This produces a sequence $x_1, \ldots, x_T$ of increasingly noisy data. The conditional model learns a reverse denoising function parameterized by a neural network ϵ_θ, which attempts to estimate the noise added at each timestep t, conditioned by simulated data c obtained by the numerical model of the ocean. x_T represents the noisy SST sample at timestep t, ϵ_θ the neural network function parameterized by θ that predicts the added noise. The learning objective minimizes the expected difference between the true and predicted noise:

$$L(\theta) = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t \mid c)\|^2\right]. \tag{1}$$

During inference, the model begins with pure Gaussian noise and iteratively applies the learned reverse process, guided by the simulated input, to generate a reconstruction that approximates the reanalysis SST. This conditional generation approach allows the model to leverage both the statistical structure learned from training data and the physical cues present in the satellite observations. We implement this model using a U-Net architecture, which is well-suited for spatial data due to its hierarchical feature extraction and reconstruction capabilities. The conditioning input is incorporated into the model either by concatenation at the input level or through adaptive feature modulation within intermediate layers, ensuring that the generative process remains guided by the simulated data throughout all stages of denoising. This makes them highly effective for oceanographic applications, where biases and uncertainty are common.

Forward Process in DDPM The forward process in diffusion models is a gradual, stochastic procedure that corrupts data by adding Gaussian noise over multiple timesteps. Inspired by non-equilibrium thermodynamics [6], this process progressively transforms a structured data sample into near-pure noise via a Markov chain. At each time-step, noise is added according to a variance schedule, with early works introducing a simple linear schedule [7]. This process can be described mathematically as a Markov process, where data x_0 is transformed into progressively noisier versions x_t through a series of steps:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I\right) \tag{2}$$

where q represents the forward process, x_t is the output of the forward process at step t (x_{t-1} is the input at step t). $\mathcal{N}$ denotes the normal distribution, $\sqrt{1-\beta_t}\, x_{t-1}$ represents the mean and $\beta_t I$ defines the variance.

Reverse Process in DDPM The reverse process aims to invert the forward degradation by iteratively denoising the sample, step by step, back to the original data distribution. A U-Net-based model was introduced that learns this denoising path by predicting the noise added in each forward step [7]. Training is performed using a simple L2 loss between predicted and actual noise. To reduce the typically slow reverse trajectory, Denoising Diffusion Implicit Models (DDIMs) were proposed, allowing for non-Markovian reverse steps that accelerate sampling without sacrificing quality [20]. Overall, the reverse process is the constructive half of diffusion models, capable of transforming random noise into high-fidelity, coherent data samples. The reverse process aims to recover the original data x_0 from a noisy sample x_t by gradually removing noise. Since the forward process is a Gaussian Markov Chain, the reverse process is also a Gaussian transition:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \tag{3}$$

Where $\mu_\theta(x_t, t)$ is the mean of the of the denoised distribution learned by the U-Net and $\Sigma_\theta(x_t, t)$ is the variance that can be learned or fixed. It basically states that p_θ (the reverse diffusion process) is a chain of Gaussian transitions starting at $p(x_t)$ and iterating T times using the Eq. 3 for one diffusion process step $p_\theta(x_{t-1} \mid x_t)$.

3 Methodology

3.1 Dataset

The dataset employed in this study was sourced from the Copernicus Marine Environment Monitoring Service (CMEMS), specifically the Mediterranean Sea Physics Reanalysis (MED-REA) product. The simulated data used to condition the diffusion model are produced by the Mediterranean forecasting system (MedFS) and kindly provided by the Euro-Mediterranean Center on Climate Change Foundation (CMCC). In this study, we considered only the sea surface temperature (SST) measurements from the Ionian Sea in southern Italy with a daily temporal resolution and spatial resolution of 1/24 degree (ca. 5 km). It captures the dynamic variations in ocean temperatures influenced by seasonal and environmental factors. For training, we use data collected over five years from 2015 to 2019, providing a diverse and comprehensive set of temperature patterns across different seasons. For testing, we select data from the year 2009, which allows us to evaluate the model's ability to generalize to unseen temporal patterns. The dataset from 2015 to 2019 was divided into 90% for training and 10% for validation. After ensuring satisfactory model performance within this period, the model was tested on 2009 data, which lies completely outside the training range. This approach was adopted to verify the temporal consistency of the model predictions. This separation between training and testing periods ensures a realistic assessment of model performance, simulating practical deployment scenarios where models encounter new environmental conditions. The SST data is structured in a consistent spatial and temporal format, making it suitable for training and evaluation without additional preprocessing steps. Overall, the dataset offers a challenging yet representative benchmark for testing the predictive capabilities of diffusion models in oceanic environments.

3.2 Model Architecture

The U-Net architecture (depicted in Fig. 1) has emerged as the de facto standard for denoising networks in diffusion models due to its powerful encoder-decoder structure and skip connections. Its hierarchical design enables both local detail preservation and global context modeling, which are crucial for generating high-quality images. The U-Net-based diffusion model architecture in our implementation consists of 4 blocks in the encoder, one bottleneck, and the mirrored decoder. Skip connections and normalization are kept to maintain image details [7]. The encoder path comprises convolutional layers, each followed by batch normalization and SiLU activation functions, progressively reducing spatial dimensions while increasing feature depth. The decoder mirrors this process, performing up-sampling via transposed convolutions and recombining information from earlier encoder layers via skip connections.

Moreover, to systematically analyze the influence of model capacity, we vary the number of channels in the convolutional layers across different experiments, testing configurations with 8, 16, 32, 64, 128 and 256 channels. This exploration helps to evaluate how architectural scaling impacts both generative quality and computational cost. Finally, the temporal embeddings described earlier are applied to each residual block of the U-Net, enabling

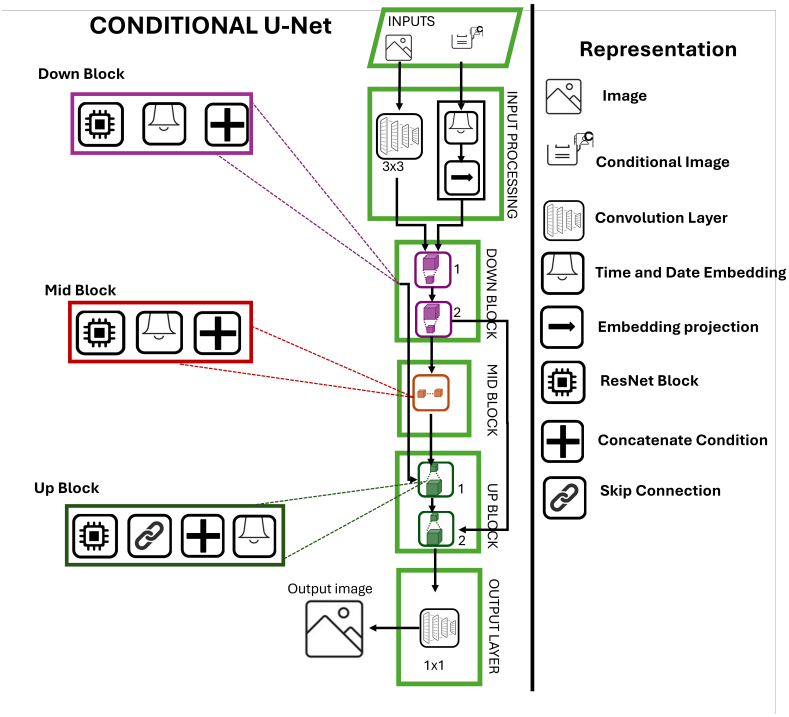

Fig. 1. Conditional U-Net architecture.

adaptive conditioning on the time-step throughout the model. The clear conditioning input state and the latent noisy representation are combined through concatenation along the channel dimension at each U-Net layer. A convolutional operation then integrates these concatenated channels, enabling effective utilization of conditioning information during reconstruction.

3.3 Training and Evaluation

In the context of oceanographic datasets, where the predicted fields (e.g., temperature, salinity, or velocity) are subject to both fine-scale variability and occasional anomalies, MAE offers a straightforward interpretation of how far the model predictions deviate from the true observations on average. Furthermore, because MAE is expressed in the same units as the predicted variable (e.g., degrees Celsius for temperature fields), it allows for direct and intuitive physical interpretation of the model's performance. Low MAE values indicate that the model's predictions closely track the true fields with small average deviations, making MAE a complementary metric to RMSE for a comprehensive evaluation of model accuracy. In addition, we used the Pearson correlation coefficient and the R^2 to assess how well the spatial patterns or variations between the two maps (the generated one and the ground truth) align each other. In the following we recap how evaluation metrics are defined.

Root Mean Squared Error (RMSE) The Root Mean Squared Error (RMSE) quantifies the average squared difference between the predicted ($\hat{y}$) and the true (y) maps:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{4}$$

where N denotes the total number of points. RMSE penalizes larger errors more heavily, making it sensitive to outliers. Minimizing the RMSE ensures that predictions closely replicate physical quantities. Moreover, it provides an interpretable measure of the model average error in the same units as the physical variables, facilitating direct comparison with observational measurements.

Mean Absolute Error (MAE) The Mean Absolute Error (MAE) provides a measure of the average magnitude of errors between predicted and true fields, without considering their direction. It is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|, \tag{5}$$

The absolute value ensures that positive and negative errors contribute equally to the final score. Unlike the Root Mean Squared Error (RMSE), which emphasizes larger errors, MAE treats all errors linearly. This property makes MAE less sensitive to outliers and provides a more robust measure of overall predictive performance, particularly in datasets where occasional large deviations may occur due to measurement noise or extreme natural events.

Pearson Correlation Coefficient (PCC) The Pearson Correlation Coefficient (PCC) measures the linear correlation between the predicted and true fields:

$$\mathrm{PCC} = \frac{\sum(\hat{y}_i - \hat{\bar{y}})(y_i - \bar{y})}{\sqrt{\sum(\hat{y}_i - \hat{\bar{y}})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{6}$$

PCC assesses whether the variations in predicted fields align linearly with those in the true fields, independently of the scale. A high PCC value indicates strong spatial agreement, essential for oceanographic modeling where large-scale gradients govern physical dynamics.

Coefficient of Determination (R^2) The Coefficient of Determination (R^2) [24] quantifies the proportion of variance in the true fields explained by the model predictions:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{7}$$

The R^2 value near 1 indicates a highly accurate model, while values near or below zero indicate poor predictive performance. For oceanographic datasets, maintaining high R^2 values ensures that the model captures both mean behavior and variability across the domain.

3.4 Model parameters

In the base configuration, the model takes one-channel input fields corresponding to the sea surface temperature. The architecture has four down-sampling stages, a mid-block, and four up-sampling stages, with a dropout regularization applied. All experiments use a fixed input resolution of 128×128 and a batch size of 8 during training. We used Adam optimizer, one of the most popular and effective optimization algorithms used in training deep learning models. Learning rate is set to $lr = 2 \cdot 10^{-5}$. The dropout rate is set to 0.1, and the batch size is set to 8. The optimizer parameters, learning rate, and training

duration are configured to ensure stable convergence.

To systematically explore the impact of model capacity and training speed, several key parameters are varied. First, the number of channels in the initial convolutional layer, referred to as the base channel depth, is varied over a range of 8, 16, 32, 64, 128, and 256. This variation tests how changing the depth of the channels influences their ability to capture multiscale spatial features critical to oceanographic phenomena. Increasing the number of base channels increases the model's representational capacity, allowing it to better approximate complex structures, while smaller channel models are lighter but may under-fit if not trained properly. In addition to base channel variations, experiments are performed to study the behavior of the model to the learning rate. Specifically, for the model with a base channel depth of 64, different learning rates are tested to evaluate how optimization behavior, including convergence speed and final accuracy, depends on this parameter. Since moderately sized models balance complexity and computational cost, tuning the learning rate is essential to achieving efficient training without instability or stagnation.

Furthermore, the impact of training duration is investigated by extending the number of training epochs for the base channel 64 model from the default setting up to 10,000 epochs. This variation aims to determine whether longer optimization helps smaller or medium-capacity models gradually learn the complex multiscale structures present in oceanographic datasets. Initial observations indicated that lower channel models benefit significantly from additional training, as they require more optimization steps to capture the variability and structure of physical fields accurately. Through these variations, the study provides a comprehensive analysis of how architectural and training choices influence the diffusion model's ability to reconstruct oceanographic fields near to the reanalysis data.

4 Results and Discussion

We initiated our analysis by examining the computed loss between the original noise and the predicted noise during the training of our conditional diffusion model, considering different base channel (B. Ch.) configurations (8, 16, 32, 64, 128, and 256) in the U-Net architecture. As shown in Figure 2, the loss decreases sharply during the initial training phase and gradually stabilizes as the number of epochs increases. This trend is observed across all base channel configurations in the U-Net architecture. Moreover, the figure shows that increasing the number of base channels in the U-Net architecture leads to a lower overall loss and accelerates the convergence process, requiring fewer epochs to reach stabilization. These findings suggest that increasing base channel configurations enhances the model's learning efficiency and improves its capacity to accurately predict the noise during training.

To assess the quality of the parameters generated by the conditional diffusion model under different base channel configurations in the U-Net architecture, we used a set of evaluation metrics. Table 1 presents the performance of the conditional diffusion model trained with different numbers of base channels [8, 16, 32, 64, 128, 256] in the U-Net architecture based on quantitative metrics, including root mean squared error (RMSE), mean absolute error (MAE), Pearson correlation coefficient (PCC), and the coefficient of determination (R^2). These metrics collectively assess the predictive accuracy, error magnitude, and correlation between the generated image i.e Biased corrected, and ground truth parameters across different base channel configurations.

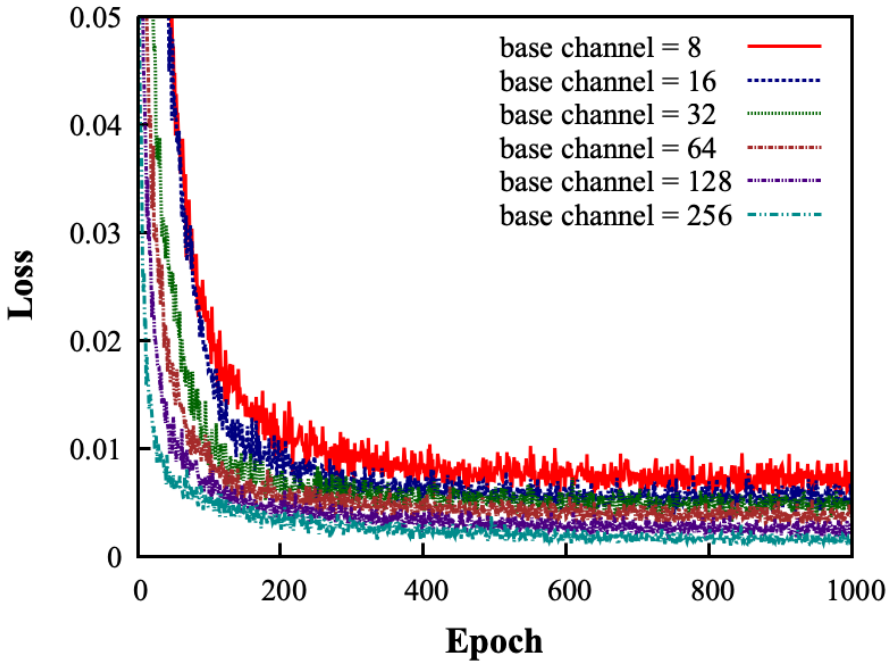

Fig. 2. Training loss vs. epochs for varying U-Net base channel depths in diffusion model noise prediction.

Table 1. Evaluation Metrics for Different Base Channel Depths

Channel	RMSE	MAE	PCC	R^2
8	1.1184	0.8504	0.9988	0.9780
16	1.1606	0.8835	0.9987	0.9761
32	0.5419	0.3729	0.9991	0.9953
64	0.4967	0.3418	0.9992	0.9957
128	0.3113	0.1967	0.9993	0.9985
256	0.2945	0.1873	0.9994	0.9986

The results show that reconstruction becomes more precise with increasing base channels. The diffusion models with smaller numbers of base channels (i.e., 8 and 16) in U-Net contain very high values of RMSE and MAE, which indicate less precise generated parameters. For example, the 8 base channel model gives an RMSE of 1.11 and an MAE of 0.85, whereas the model with 256 base channels gives comparatively smaller errors (RMSE = 0.29, MAE = 0.18), demonstrating the benefit of using models of higher capacity. In addition, PCC and R^2 provide a clear trend of improvement against a higher number of channels. The Pearson Correlation Coefficient (PCC) grows from 0.9988 for the 8-channel model to 0.9994 for the 256-channel model, showing a greater linear relationship between actual and predicted values. The coefficient of determination (R^2) also improves from 0.9780 to 0.9986, implying that the 256-channel model explains nearly all the variation in the data.

Notably, a key inflection point in performance is observed between 32 and 64 channels. Table 2 highlights this transition in more detail. While the reduction in error metrics from 32 to 64 base channels is moderate, the overall trend confirms a meaningful gain in precision. The PCC and R^2 values already approach saturation within this range, with only marginal improvements observed thereafter. Nonetheless, the increase from 32 to 64 channels enables the model to better capture fine-grained spatial details and complex patterns,

Table 2. Performance comparison between 32 and 64 base channel depths in U-Net architecture.

Metric	B.Ch. 32	B.Ch. 64	Relative Improvement
RMSE	0.5419	0.4967	8%
MAE	0.3730	0.3418	8%
PCC	0.9991	0.9992	Slight
R^2	0.9953	0.9958	Slight

which may be critical for downstream tasks. Figure 3 visually illustrates the performance of the conditional diffusion model in reconstructing oceanic temperature fields across different base channel configurations in the U-Net architecture. Each row corresponds to a model trained with a specific base channel depth, allowing for a direct visual comparison of how channel depth influences reconstruction accuracy. Across all rows, the biased corrected temperature fields (third column) closely resemble the reanalysis data (ground truth) (first column), indicating the model's overall effectiveness in learning the underlying data distribution. The conditioning input (second column), derived from simulation data, appears to guide the model toward realistic and physically consistent outputs. A key observation emerges from the difference maps (fourth column). Models with fewer base channels (e.g., 8 or 16) exhibit visibly higher reconstruction errors, as indicated by the more intense red regions in the error maps. These differences suggest that models with low channel capacities struggle to capture finer spatial features and temperature gradients.

As the number of base channels increases (e.g., from 32 to 256), the error maps become progressively lighter and more diffuse, reflecting lower reconstruction errors and improved alignment with the ground truth. The visual representation shows an improvement with increasing base channel and is consistent with the quantitative metrics reported in Table 1. We observed the reductions in RMSE and MAE with increasing base channel depth. We also observed a distinct improvement between the 32- and 64-channel configurations. The model with 64 base channels captures sharper boundary features and better preserves spatial coherence, which aligns with the performance inflection point identified in the quantitative analysis. The corresponding difference map for this configuration is significantly lighter, particularly in coastal and high-gradient regions, confirming that increased representational capacity enhances the model's ability to reconstruct complex spatial structures. However, further increasing the base channels to 128 and 256 yields only marginal visual improvements. This suggests diminishing returns, where qualitative gains in inference quality become less significant despite continued improvements in evaluation metrics. These findings reinforce the trade-off between model complexity and computational cost discussed earlier. Thus, the visual analysis supports that increasing the number of base channels enhances the fidelity of the biased corrected temperature fields. The 64-channel configuration represents an effective balance, offering high-quality reconstructions with relatively low reconstruction error while avoiding excessive computational overhead.

Table 3 summarizes the size and complexity of the conditional diffusion models trained with different base channel sizes. As the number of base channels increases, the number of trainable parameters grows rapidly. Specifically, models with 8 and 16 base channels maintain relatively lightweight architectures, with only 427 thousand and 1.3 million parameters, respectively. These models require minimal memory (1.7–5.2 MB) and are well-suited for resource-constrained environments. In contrast, models with 32 and 64 channels significantly increase in complexity, reaching 4.6 million and 17.2 million parameters, respectively, corresponding to memory footprints of approximately 18 MB and 69 MB. These configurations offer a good balance between model expressiveness and computational effi-

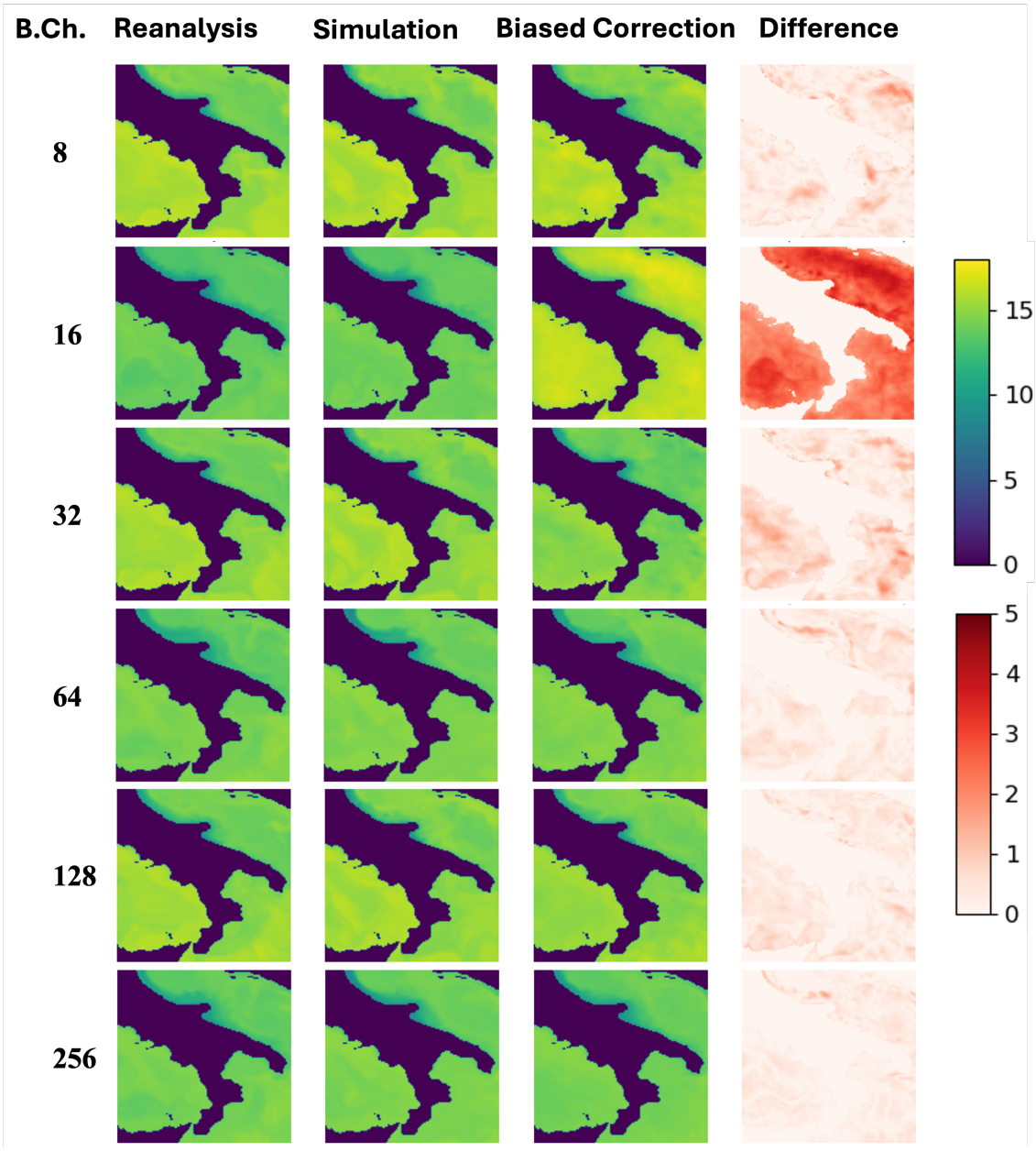

Fig. 3. Sea surface temperature (SST, °C) inference using a conditional diffusion model varying the U-Net depths. Rows: base channel configurations. Columns: reanalysis (ground-truth), input (conditional), bias-corrected output, and absolute error (spatial reconstruction differences).

ciency, making them attractive for practical deployment. At the highest capacities, models with 128 and 256 base channels show a dramatic rise in parameter counts to 67.2 million and 265 million, respectively, with memory requirements of approximately 269 MB and over 1 GB. The "Total Estimated Model Parameters Size (MB)" metric refers to the approximate amount of memory needed to store all trainable parameters in memory during training and inference. It accounts for the storage of each parameter (typically as 32-bit floating-point values) and provides a realistic estimate of the model's memory footprint. This metric is critical for assessing the feasibility of training models on available hardware, particularly when working with high-resolution scientific data. Thus, while increasing base channel size improves model capacity and reconstruction quality, it also leads to substantial growth in memory consumption and computational cost. Selecting an optimal configuration requires balancing these trade-offs based on specific application constraints and available computational resources.

Table 3. Relationship between the U-Net base channel sizes, the number of trainable parameters, and the total estimated model parameter size.

Base Ch	Trainable Parameters	Tot. Est. Model Param. Size (MB)
8	427 K	1.710
16	1.3 M	5.184
32	4.6 M	18.268
64	17.2 M	68.988
128	67.2 M	268.634
256	265 M	1060.749

Importantly, this 64 base channel configuration offers a favorable trade-off between accuracy and computational cost. While larger channel depths such as 128 and 256 continue to yield performance gains, these improvements come with diminishing returns relative to their resource demands. Thus, increasing the base channel capacity enhances both the expressiveness and precision of the model. However, practical considerations such as memory consumption and inference speed must be carefully weighed when selecting the optimal configuration for deployment.

To further assess the model's temporal generalization capability, we analyzed its performance in generating sea surface temperature across different seasonal batches. Each batch consisted of two consecutive months, and for each batch, 30 random test images were used to compute the evaluation metrics. The model used for this analysis was trained with a base channel depth of 64. The batch-wise performance results are summarized in Table 4.

Table 4. Bimonthly performance metrics diffusion model in predicting sea surface temperature.

Batch	RMSE	MAE	PCC	R^2
Jan - Feb	0.4617	0.3172	0.9989	0.9945
Mar - Apr	0.5497	0.4062	0.9993	0.9920
May - Jun	0.3472	0.2324	0.9996	0.9985
Jul - Aug	0.5309	0.3826	0.9997	0.9976
Sep - Oct	0.5011	0.3351	0.9994	0.9977
Nov - Dec	0.5893	0.3772	0.9983	0.9940

Overall, in all seasons, the model showed robust generative performance, with Pearson's

correlation coefficients of more than 0.998 and the coefficient of determination (R^2) of more than 0.99, reflecting outstanding agreement between the biased corrected and reanalysis SST (ground truth). Minimum error measures (RMSE = 0.3472, MAE = 0.2324) were recorded in May-June, indicating that the model performed optimally in late spring through early summer. In contrast, slightly higher errors occurred for November-December and March-April due to greater variability in SST or atmospheric influences during times of seasonal change. In addition to the quantitative evaluation, a further qualitative check was conducted by visual inspection of the generated SST fields. Figure 4 presents a grid layout of representative samples from each period. The top row displays the reanalysis data (ground truth SST), the second row shows the simulation data, the third row illustrates the SST fields generated by the model, and the bottom row presents the absolute error maps, highlighting the pixel-wise differences between the biased corrected and ground truth SST. The SST maps generated exhibit strong visual agreement with the reanalysis data, effectively capturing both the spatial structures and the temperature gradients. The error maps, predominantly light in color, indicate low residual differences in the domain, with slightly higher errors concentrated in regions characterized by sharp gradients or complex coastal features. This visual consistency further supports the model's ability to produce realistic and high-fidelity SST reconstructions across different seasonal periods.

To qualitatively assess the performance of the diffusion model in generating realistic sea surface temperature (SST) fields, we conducted a visual analysis across different training stages (2000, 4000, 6000, 8000, and 10,000 epochs), with a fixed base channel depth of 64. The visual comparison is illustrated in Figure 5. As training progresses, the quality of the generated SST fields improves markedly. The absolute error maps in the bottom row of Figure 5 highlight regions of significant pixel-wise deviation between biased corrected and ground truth SST. At earlier epochs (2000, 4000), large areas of elevated error are apparent, particularly in regions characterized by complex thermal dynamics. These discrepancies diminish progressively with training, with the model between 8000 and 10,000 epochs exhibiting the lowest error magnitude and spatial extent. Notably, the remaining residual errors are primarily concentrated in zones of sharp thermal gradients, such as oceanic fronts, suggesting that while the model learns broader SST patterns well, capturing abrupt transitions remains more challenging. To complement the visual analysis, we quantitatively evaluated the SST outputs against the reanalysis data using standard metrics. Table 5 summarizes the performance across epochs.

Table 5. Evaluation metrics for bias-corrected SST images against reanalysis data at different training epochs (base channel depth = 64).

Epoch	RMSE	MAE	PCC	R^2
2000	0.29887	0.18909	0.99941	0.99867
4000	0.29132	0.18257	0.99943	0.99875
6000	0.28425	0.17819	0.99946	0.99880
8000	0.27890	0.17448	0.99946	0.99883
10000	0.28110	0.17566	0.99946	0.99882

The metrics confirm the trend in visual analysis. Specifically, RMSE decreases from 0.298 for 2000 epochs to 0.278 for 8000 epochs, accompanied by corresponding improvements in MAE. PCC remains always very high (greater than 0.9994), indicating a strong linear relationship between the biased corrected and reanalysis (ground-truth) SST values, whereas

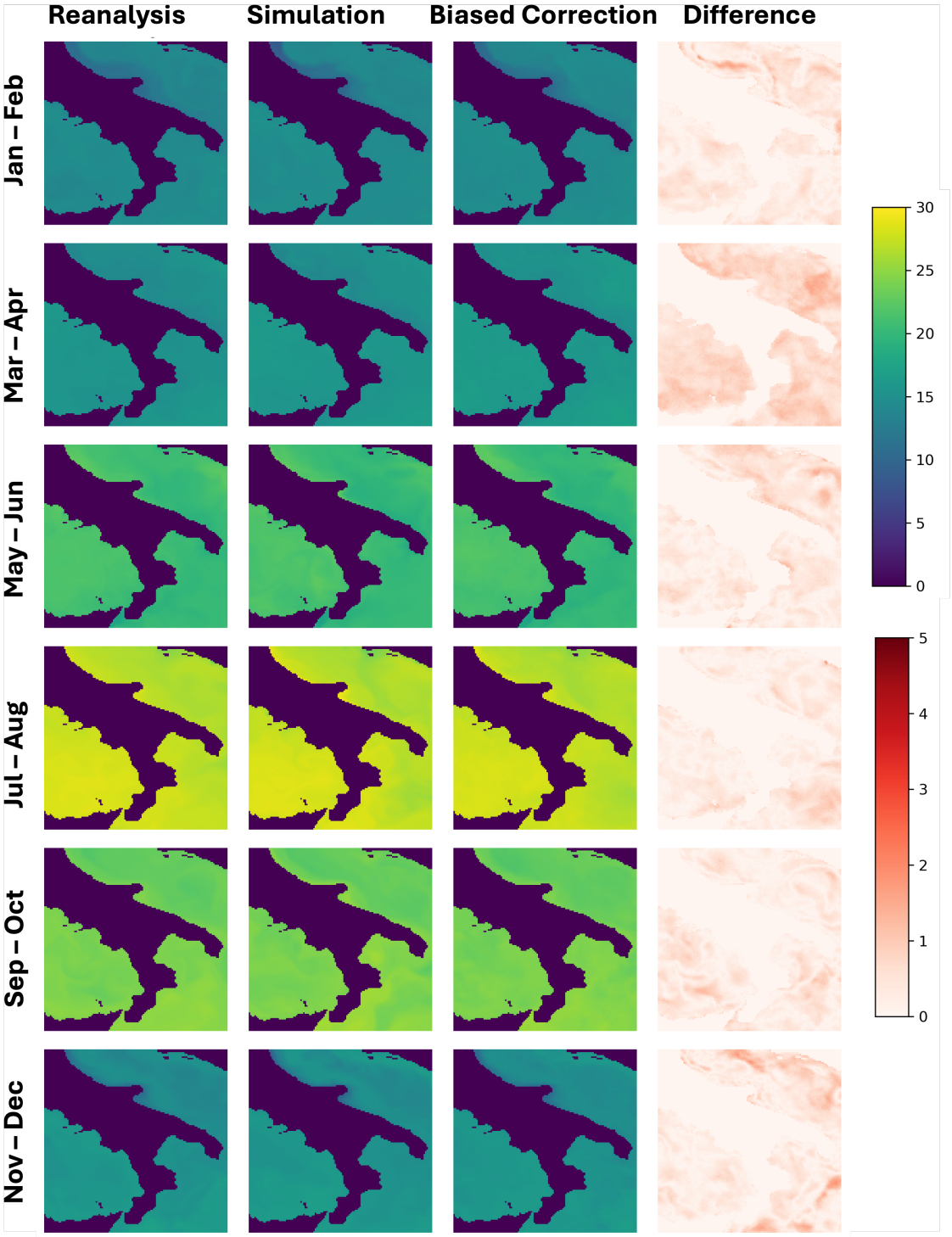

Fig. 4. Qualitative analysis of sea surface temperature (SST, °C) predictions for selected samples across different bimonthly batches.

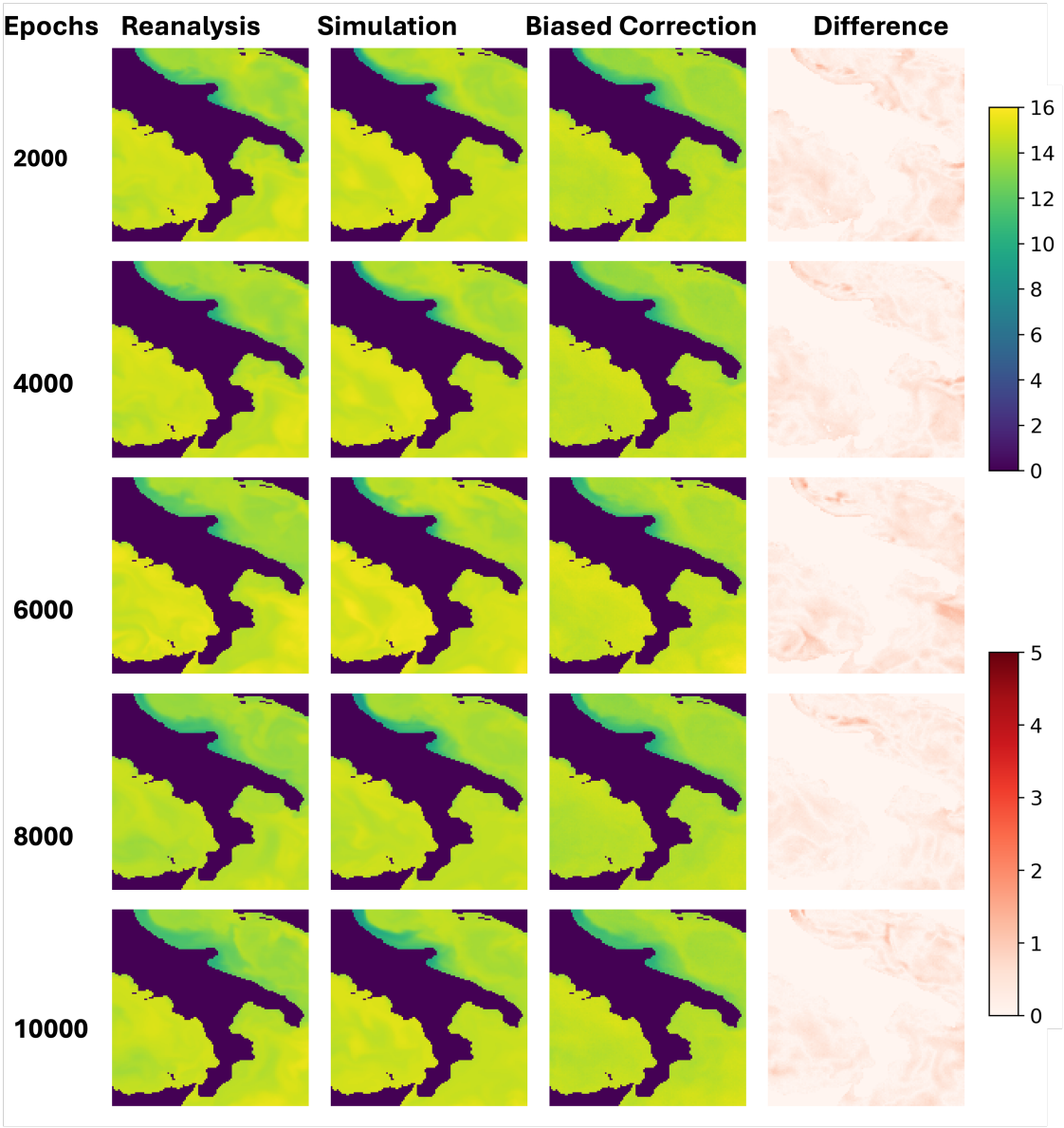

Fig. 5. Qualitative evaluation of biased correction of sea surface temperature (SST, °C) fields by the diffusion model with base channel depth 64 across different training epochs (2000, 4000, 6000, 8000, and 10000).

R^2 values confirm high explanatory power at all considered epochs. Interestingly, a performance drop is observed at 10,000 epochs from 8,000, suggesting overfitting or plateauing in model training. Thus, the results indicate that the diffusion model learns to reconstruct SST fields with greater accuracy as training progresses. Although further training for over 8000 epochs yields marginal improvement, the model shows good spatial agreement and correlation with actual SST observations to establish its validity and practicality for actual oceanographic use. Particularly, the model using a base channel depth of 64, trained for 8000 epochs, equals the model using 256 base channels, which was trained for 1000 epochs. These findings establish a trade-off between model capacity (i.e., base channel depth) and training time (i.e., number of epochs). With sufficient training, a fairly sized model (base channel depth 64) will be able to match, and in some cases outperform, the performance of larger models (base channel 128 or 256) given less training time (1000 epochs). This reveals that longer training can successfully override models' base channel depth, providing an alternative solution more computationally intensive without compromising prediction performance.

To investigate the effect of learning rate on the training stability and performance of the diffusion model, we conducted a series of experiments by training the diffusion model with a base channel of 64 and five different learning rate: $2 \cdot 10^{-2}$, $2 \cdot 10^{-3}$, $2 \cdot 10^{-4}$, $2 \cdot 10^{-5}$, and $2 \cdot 10^{-6}$. The training loss was monitored over 1000 epochs, and the results are summarized in Figure 6

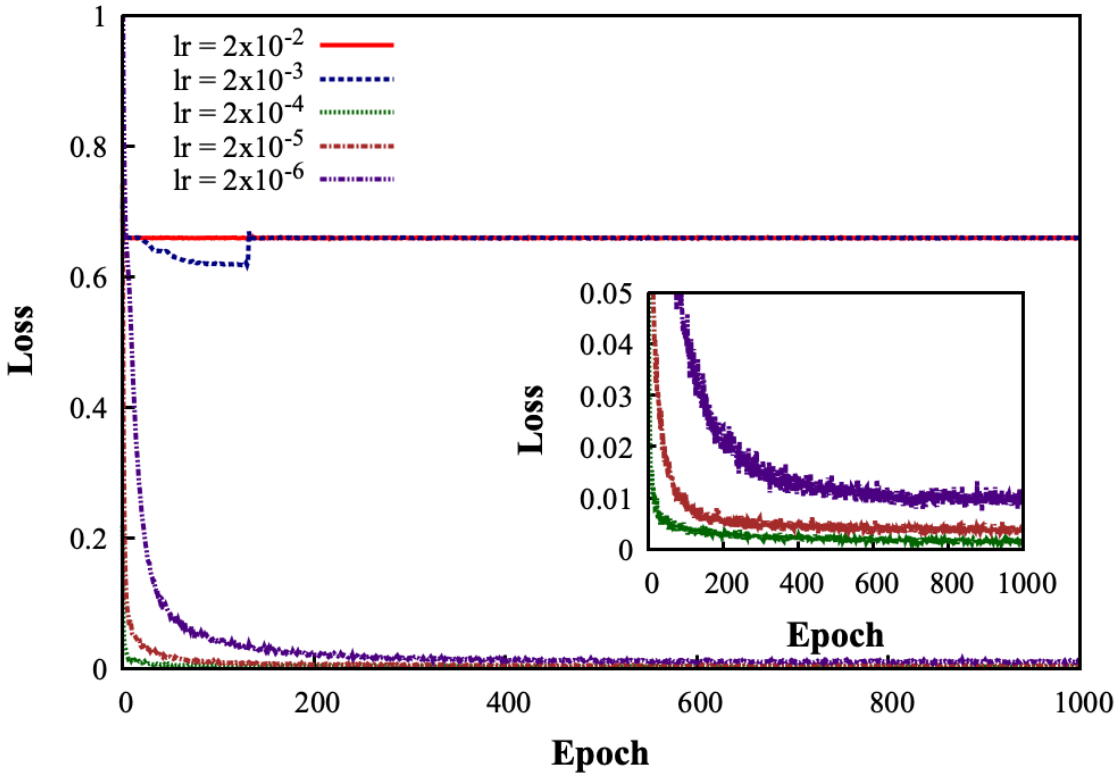

Fig. 6. Training loss curves of the diffusion model (base channel = 64) trained with different learning rates over 1000 epochs.

It can be seen from Figure 6 that the learning rate plays an important role in affecting convergence behavior and final performance of the model. The largest learning rates, $2 \cdot 10^{-2}$, and $2 \cdot 10^{-3}$, resulted in poor convergence with the loss being stuck at rather high values (≈ 0.68 and ≈ 0.66 respectively). It suggests that such learning rates are too large, potentially causing the optimizer to overshoot the minima. On the other hand, the learning rate of $2 \cdot 10^{-4}$ produced fast and stable convergence with the minimum final loss of approximately 0.0025. This suggests that $2 \cdot 10^{-4}$ is the optimal learning rate among those that were attempted, regarding both learning speed and stability. A lower learning rate of $2 \cdot 10^{-5}$ also showed convergent stability but slower and at a marginally higher end loss (≈ 0.004), and the lowest learning rate $2 \cdot 10^{-6}$ showed extremely slow convergence and plateaued at a relatively higher loss (≈ 0.008), reflecting its inefficiency. A zoomed-in plot is added to the figure for a more direct comparison of the loss behavior for smaller learning rates, where the curves are nearly on top of one another in the main plot. This zoomed view makes the superior performance of $2 \cdot 10^{-4}$ and its advantage over others clearer. These findings underscore the importance of selecting an appropriate learning rate for training diffusion models and demonstrate that $2 \cdot 10^{-4}$ yields the best trade-off between convergence speed and final accuracy in this configuration.

5 Conclusion

This study gives a detailed evaluation of a conditional diffusion model used for bias correction of the Sea Surface Temperature (SST) fields, with particular interest in the base channel depth contribution towards the U-Net backbone network. The results of the experiments confirm that the number of base channels greatly influences the ability of the model to learn, converge, and generalize. Loss analysis during training shows a similar pattern of rapid drop and convergence in all base channel configurations, with larger base channel depth in the U-Net architecture converging more quickly and having smaller values of final loss. Quantitative measures of performance like RMSE, MAE, PCC, and R^2 show a similar and clear improvement in prediction accuracy with an increase in the number of base channels. Specifically, the 8 or 16-channel models have significantly higher error rates, while the 128 or 256-channel models have excellent accuracy, albeit with diminishing marginal returns beyond 64 channels. The 64-channel model is a point of inflection, offering the best trade-off between computation efficiency and reconstruction quality. Visual inspection also supports these quantitative results. Lower base channel depth, model-generated reconstructed SST fields have more conspicuous spatial errors, particularly in areas of high gradient or coastal complexities. Conversely, higher base channel depth creates visually coherent and thermodynamically reasonable SST maps, and the error maps support higher spatial fidelity. Of special note is the model with 64 base channels, which has phenomenal accuracy in reproducing large-scale and fine-scale oceanographic structures.

Temporal generalization tests across a bimonthly time frame underscore the model's robustness. High correlation metrics and low reconstruction errors across diverse periods suggest that the model captures both persistent and transient SST patterns effectively. Visual inspection of seasonal SST predictions corroborates these results, with the model consistently generating outputs that closely align with reanalysis data. Minor discrepancies, typically confined to zones with abrupt SST transitions, indicate areas where further refinement may be necessary, perhaps through specialized attention mechanisms or higher-resolution inputs. Furthermore, analysis of model output across training epochs reveals a gradual but consistent enhancement in biased corrected SST field quality. By epoch 8000 and beyond, the reconstructions become nearly indistinguishable from ground truth data,

signaling model convergence and maturity in learning complex spatial structures.

In conclusion, this study provides evidence that increasing base channel depth improves the expressiveness and accuracy of conditional diffusion models for SST prediction. However, beyond a certain threshold, identified as 64 channels, the gains in performance become marginal relative to the increase in computational cost. These insights can guide future work in deploying diffusion-based generative models for ocean modeling, weather forecasting, and climate simulations, especially where computational efficiency is a priority. Future extensions could explore multi-variable conditioning, 3D spatial reconstructions, and real-time ocean monitoring systems integration. Accurately capturing fine-scale spatial features, particularly in coastal regions, is critical for real-world ocean forecasting and marine resource management. Coastal zones often exhibit sharp temperature gradients and are strongly influenced by localized currents, freshwater inflows, and topographic complexity. Misrepresenting these features can lead to significant errors in downstream applications such as habitat modeling, coastal upwelling prediction, or marine heatwave detection. Therefore, improving reconstruction fidelity in these areas enhances the reliability of high-resolution ocean models used for operational forecasting and decision-making. While the current model demonstrates strong performance in reconstructing sea surface temperature (SST) fields over the Mediterranean Sea, it is not inherently limited to this region. Our choice was guided by the goal of developing diffusion models for oceanographic data assimilation, with the Mediterranean Sea serving as an initial testbed. The model architecture is scalable and can be extended to larger datasets, higher-resolution fields, and 3D variables. Our future work aims to incorporate deeper ocean temperatures, additional variables such as salinity and velocity, and applications beyond the Mediterranean Sea region.

References

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets, Advances in Neural Information Processing Systems 27 (2014) 2672-2680.
2. D.P. Kingma, and J. Ba. Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980 (2014).
3. A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel Recurrent Neural Networks, arXiv preprint arXiv: 1601.06759 (2016).
4. L. Dinh, R. Pascanu, S. Bengio, Y. Bengio. Sharp Minima Can Generalize For Deep Nets, arXiv preprint arXiv: 1703.04933 (2017).
5. D. P. Kingma and P. Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions, arXiv preprint arXiv: 1807.03039 (2018).
6. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning (ICML), July 2015, pp. 2256-2265.
7. J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models, Advances in Neural Information Processing Systems 33 (2020) 6840-6851.
8. A. Brock, J. Donahue, and K. Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis, arXiv preprint arXiv: 1809.11096 (2019).
9. A. Razavi, A. van den Oord, and O. Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2, arXiv preprint arXiv: 1906.00446 (2019).
10. P. Esser, R. Rombach, and B. Ommer. Taming Transformers for High-Resolution Image Synthesis, arXiv preprint arXiv: 2012.09841 (2021).
11. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 105-114.
12. S. Bell-Kligler, A. Shocher, and M. Irani. Blind Super-Resolution Kernel Estimation using an Internal-GAN, arXiv preprint arXiv: 1909.06581 (2020).

13. C. Saharia, J. Ho, W. Chan, T. Salimans, D.J. Fleet, and M. Norouzi. Image Super-Resolution via Iterative Refinement, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (4) (2023) 4713-4726.
14. T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y, Zhu. Semantic Image Synthesis With Spatially-Adaptive Normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019, pp. 2332-2341.
15. X. Huang, A. Mallya, T.-C. Wang, M.-Y. Liu. Multimodal Conditional Image Synthesis with Product-of-Experts GANs. In Proceedings of the Computer Vision – ECCV, October 2022, pp. 91-109.
16. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. Kamyar, S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. Gontijo Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, arXiv preprint arXiv: 2205.11487 (2022).
17. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv preprint arXiv: 2204.06125 (2022).
18. A. Nichol, and P. Dhariwal. Improved Denoising Diffusion Probabilistic Models, arXiv preprint arXiv:2102.09672 (2021).
19. Y. Song, and S. Ermon. Score-Based Generative Modeling through Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations, April 2020.
20. Y. Song, C. Meng, and S. Ermon. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations, May 2021.
21. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2022, pp. 10684-10695.
22. J. Ho, and T. Salimans. Classifier-Free Diffusion Guidance, arXiv preprint arXiv:2207.12598 (2022).
23. O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv preprint arXiv: 1505.04597 (2015).
24. I. Goodfellow, Y. Bengio, A. Courville. Deep Learning, MIT Press (2016).

Acknowledgements

This work was partially funded under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.4 – Call for Tender No. 1031 of 17/06/2022 by the Italian Ministry for University and Research, funded by the European Union – NextGenerationEU (Project No. CN-00000013). RS acknowledges the CINECA award under the ISCRA initiative for the provision of high-performance computing resources and technical support.

Authors

M. Sarmad received an M.S. in Electrical and Electronics Engineering from Istanbul Aydin University, and he did a Bachelor's degree in Electrical Engineering. Currently, he is pursuing his PhD in the Department of Engineering for Innovation at, University of Salento. His research interests include artificial intelligence, machine learning.

E. Mele received his MSc degree in Computer Engineering in 2024 from the Department of Engineering, University of Salento, and is currently a PhD candidate at the same university. His research interests are in deep learning, image processing, and statistical algorithms.

R. Srivastava received a Ph.D. degree in Physics from CSJM University, Kanpur, India. He is currently a Researcher at the University of Salento with over a decade of experience in multiscale and molecular modeling. His work spans high-performance computing, algorithm development, and AI-driven simulations within several European research projects.

M. Pulimeno received the Ph.D. degree in mathematics and computer science from the Department of Engineering for Innovation, University of Salento. He is currently a Researcher with the Department of Engineering for Innovation, University of Salento. His research interests include streaming algorithms, distributed and high-performance computing, data mining, and machine learning.

M. Cafaro (Senior Member, IEEE) received the Laurea degree (M.Sc.) in computer science from the University of Salerno, and the Ph.D. degree in computer science from the University of Bari. He is currently an Associate Professor with the Department of Engineering for Innovation, University of Salento. He is the Director of the Master's in Applied Artificial Intelligence and the Head of the HPC Laboratory, University of Salento.

I. Epicoco received the Ph.D. degree in innovative materials and technologies from ISUFI, University of Lecce, Italy, in 2003. He is currently an Associate Professor with the University of Salento, Lecce. His research interests include data mining, machine learning, and parallel algorithms. He is also principal scientist at the Euro-Mediterranean Center on Climate Change Foundation (CMCC).