



MAESSTRO: Masked Autoencoders for Sea Surface Temperature Reconstruction under Occlusion

Edwin Goh, Alice Yepremyan, Jinbo Wang, and Brian Wilson

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91011, USA

Correspondence: Jinbo Wang (jinbo.wang@jpl.nasa.gov)

Received: 23 June 2023 – Discussion started: 18 July 2023

Revised: 27 June 2024 – Accepted: 1 July 2024 – Published: 24 October 2024

Abstract. This study investigates the use of a masked autoencoder (MAE) to address the challenge of filling gaps in high-resolution (1 km) sea surface temperature (SST) fields caused by cloud cover, which often result in gaps in the SST data and/or blurry imagery in blended SST products. Our study demonstrates that MAE, a deep learning model, can efficiently learn the anisotropic nature of small-scale ocean fronts from numerical simulations and reconstruct the artificially masked SST images. The MAE model is trained and evaluated on synthetic SST fields and tested on real satellite SST data from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor on the Suomi NPP satellite. We demonstrate that the MAE model trained on numerical simulations can provide a computationally efficient alternative for filling gaps in satellite SST. MAE can reconstruct randomly occluded images with a root mean square error (RMSE) of under 0.2 °C for masking ratios of up to 80 %. A trained MAE model in inference mode is exceptionally efficient, requiring 3 orders of magnitude (approximately 5000×) less time compared to the conventional approaches of cubic radial basis interpolation and Kriging tested on a single CPU. The ability to reconstruct high-resolution SST fields under cloud cover has important implications for understanding and predicting global and regional climates and detecting small-scale SST fronts that play a crucial role in the exchange of heat, carbon, and nutrients between the ocean surface and deeper layers. Our findings highlight the potential of deep learning models such as MAE to improve the accuracy and resolution of SST data at kilometer scales. This presents a promising avenue for future research in the field of small-scale ocean remote sensing analyses.

1 Introduction

Submesoscale ocean fronts are regions in the ocean where two water masses with different temperatures and salinities meet, creating a boundary between them, forming a strong horizontal buoyancy gradient. These fronts are often generated by large-scale strain fields or surface buoyancy/wind forcing (Thomas, 2008; Fox-Kemper et al., 2008; McWilliams, 2016) and associated with strong vertical velocity (Mahadevan and Tandon, 2006). They can influence the exchange of heat (Su et al., 2018), carbon, and nutrients between the ocean surface and deeper layers (Mahadevan, 2016; Lévy et al., 2018), as well as the air–sea heat fluxes (Seo et al., 2023). They also function as ecotones and form rich structures in marine ecosystems. Ocean fronts are also important for human activities such as fishing, as many commercially important fish species are often found near ocean fronts (Woodson and Litvin, 2015; Belkin, 2021; Prants, 2022), where the nutrient-rich waters provide abundant food sources. As such, identification and detection of the oceanic submesoscale frontal features have a wide variety of important applications.

Detection of ocean fronts is mostly based on satellite sea surface temperature (SST) and/or ocean color. In the 1980s, the launch of the Advanced Very High Resolution Radiometer (AVHRR) enabled the detection of ocean fronts in sea surface temperature images. In the 1990s, the launch of the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) provided additional capabilities for detecting ocean fronts. Today, in addition to traditional sensors, such as AVHRR, new satellites with advanced capabilities, such as the Visible Infrared Imaging Radiometer Suite (VIIRS), continue to improve the accuracy and resolution of SST images. However, while the spa-

tial pixel resolution of VIIRS SST can reach 750 m, it uses the visible band and is thus susceptible to cloud cover.

Reconstructing SST due to cloud occlusions is an active research area due to certain limitations of satellite-based SST measurements. Ćatipović et al. (2023) extensively surveyed this topic, including reconstructions of other ocean variables such as surface chlorophyll-*a* concentration and sea surface salinity. Most operational SST products employ optimal interpolation (OI) (Jung et al., 2022), a technique that interpolates missing values based on correlations between known SST points. However, OI struggles to resolve features smaller than 100 km, resulting in overly smooth products (Chin et al., 2017). Empirical orthogonal function (EOF) methods, such as Data Interpolating Empirical Orthogonal Functions (DINEOF) (Alvera-Azcárate et al., 2011), use linear decomposition of SST for interpolation. These methods are popular but, like OI, they tend to smooth out smaller features due to truncation errors (Fablet et al., 2017; Barth et al., 2020).

Both OI-based and EOF-based methods come with assumptions of linearity and data sparsity, which often do not hold true for sparse, cloud-obstructed, and highly nonlinear SST fields (Jung et al., 2022). Hence, the focus is shifting to machine learning techniques that can better handle nonlinear dynamics. For instance, Data Interpolating Convolutional AutoEncoder (DINCAE) (Barth et al., 2020) uses neural networks for interpolation and has been successfully applied to reconstruct chlorophyll-*a* concentrations and SST (Han et al., 2020; Jung et al., 2022; Barth et al., 2022). However, compared to convolutional approaches, vision transformers used in masked autoencoders (MAEs) (He et al., 2022) have been shown to better integrate global information (Trockman and Kolter, 2022).

In this paper, we conducted a proof-of-concept test on MAE using simulated SSTs with a focus on the performance at about 10 km frontal scales, which is largely absent in existing studies. This approach, adopted from He et al. (2022), trains an encoder network to extract a compact representation of *masked* input data and simultaneously trains a decoder to reconstruct the original input from the masked representation. MAE was designed for visual representation learning to enable downstream tasks such as image classification and segmentation. In this context, the ability to reconstruct masked portions of an image presents a valuable secondary benefit. However, it is this ability that makes MAE a promising candidate for reconstructing high-resolution SST fields under cloud cover. While deep neural networks have been shown to be able to perform gap filling (e.g., Krasnopolsky et al., 2016), this is the first exploration of applying MAE in the reconstruction of high-resolution SST under cloud cover with a focus on highly anisotropic frontal structures. The modified MAE for SST Reconstruction under Occlusion is herein referred to as MAESSTRO.

This paper is organized as follows: Sect. 2.1 expounds upon the MAE methodology, followed by Sect. 2.2, which elaborates on the data used for training and validation. Sec-

tions 2.5 and 5 present the validation results for the global simulated data and a high-resolution satellite VIIRS SST snapshot at 750 m pixel resolution, respectively. Section 6 addresses the limitations of the current MAESSTRO implementation and outlines future research directions. The study concludes with a summary in Sect. 7.

2 Methodology

2.1 Masked autoencoders

MAESSTRO adapts the masked autoencoder (MAE) from natural RGB images to single-channel SST fields. The process of training state-of-the-art computer vision models requires a large number of labeled images, the acquisition of which can be prohibitively expensive (Goh et al., 2022). In contrast, natural language processing (NLP) has successfully addressed this challenge through self-supervised pretraining on unlabeled data, resulting in remarkable achievements such as the transformer-based language model Generative Pre-trained Transformer (GPT) (Radford et al., 2018).

Inspired by this progress, He et al. (2022) adapted masked autoencoding, a self-supervised learning technique rooted in language modeling (Peters et al., 2018), to computer vision, proposing MAE for self-supervised pretraining on unlabeled images. The MAE implementation for images deviates from NLP in several aspects, including the utilization of vision transformers (ViTs) (Dosovitskiy et al., 2020) to better encode positional and mask tokens, masking a larger portion of the image to account for spatial redundancy of pixels in natural images, and employing more sophisticated decoders to handle the absence of semantic information in pixels.

Figure 1 shows an example of the MAESSTRO architecture derived from the original MAE open-source code (<https://github.com/facebookresearch/mae>, last access: 1 March 2023) by He et al. (2022). During training, a random portion of SST patches is masked/removed, and the encoder only processes the unmasked patches (thereby increasing computational performance). Following this, a set of learnable mask tokens is integrated into the encoded patches, before being processed by the decoder. Ultimately, the decoder “predicts” the original SST tile in pixel format, achieving SST reconstruction through the MAE framework.

An extensive hyperparameter search resulted in MAESSTRO using an MAE variant with a patch size of 4 and a ViT-tiny encoder (Dosovitskiy et al., 2020). Here, “patch 4” refers to the process of dividing the input image into non-overlapping patches of size 4×4 pixels. The ViT model is designed to handle images as sequences of patches, much like a language model processes sequences of words or tokens. By breaking down the image into 4×4 pixel patches, the ViT encoder can effectively process and analyze the spatial structure and features of the image. This approach allows the ViT to leverage the advantages

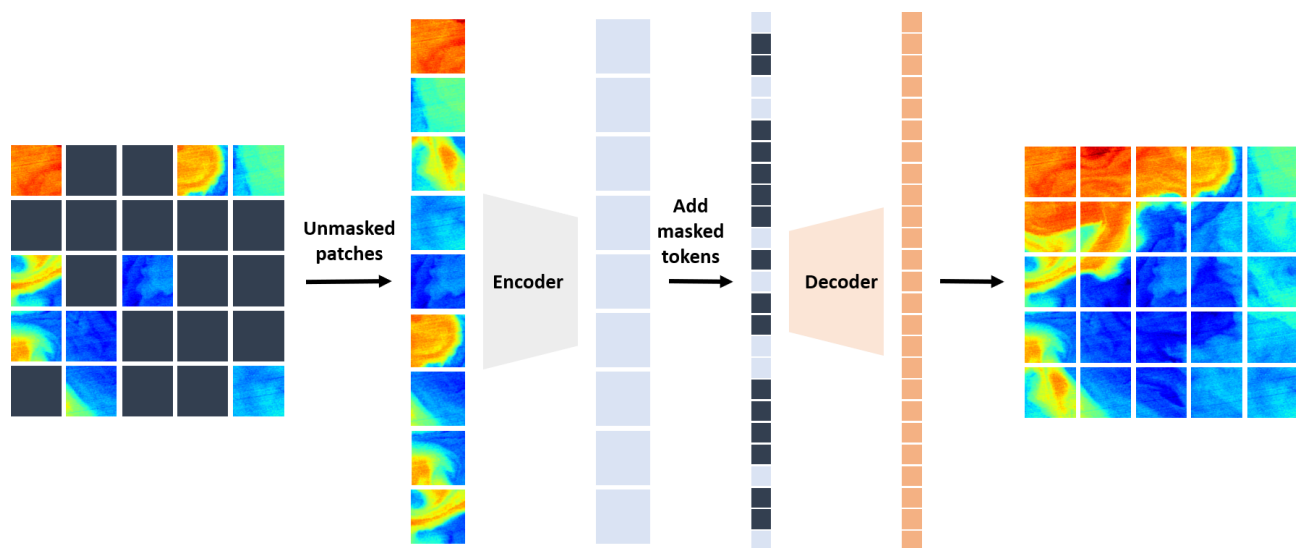


Figure 1. MAE architecture applied to SST reconstruction. When training the MAE, a random subset of SST patches is masked out. To increase computational efficiency, only the visible patches are fed into the encoder. A learnable set of *mask tokens* are added into the encoded patches and processed by the decoder, which reconstructs the original SST tile in pixels.

of the transformer architecture in computer vision tasks. In this proof-of-concept study, we exclude patches if any of the 4×4 pixels contain missing values.

2.2 Synthetic SST from high-resolution ocean simulations LLC4320 and LLC2160

To build a model for SST, using real satellite SST imagery as ground truth would be ideal. However, these images often contain noise and are susceptible to bias and errors. As an initial conceptual demonstration, this paper employs synthetic satellite SST data derived from two high-resolution numerical simulations: LLC4320 and LLC2160. These simulations were generated using the Massachusetts Institute of Technology general circulation model (MITgcm) with the same model configuration but different grid spacings. Specifically, LLC4320 utilizes $1/48^\circ$ grid spacing on a lat–lon–cap (LLC) grid, while LLC2160 uses $1/24^\circ$ grid spacing. The higher-resolution LLC4320 simulation is used as the training set, while the lower-resolution LLC2160 simulation serves as the unseen test set to evaluate MAESSTRO’s ability to generalize to unseen data of a different spatial resolution. These fields do not have added noise and errors but can contain deviations from reality due to model imperfections.

The LLC grid consists of 13 tiles covering the global ocean. Each tile of LLC4320/2160 contains 4320×4320 (2160×2160) grid points, leading to the nomenclature LLC4320/LLC2160. The simulation is forced by atmospheric input from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim reanalysis, including barotropic tides and river runoff. The LLC4320 simulation used to train MAESSTRO produces hourly snapshots

of the ocean state, spanning approximately 14 months from 13 September 2011. LLC2160 has a lower spatial resolution and was used to test the trained model at a global scale. Due to the high resolution, global coverage, and high-frequency output, the LLC4320 model is widely employed in small-scale oceanographic research (e.g., Wang et al., 2019; Su et al., 2018; Dong et al., 2021).

It is crucial to note that the foregoing are forced model simulations; thus we do not expect them to correspond exactly to the real ocean. Nevertheless, they can generate realistic small-scale ocean features that are valuable for training MAESSTRO. The absence of cloud cover in this dataset (i.e., complete SST visibility) enables its use as ground truth when evaluating the MAE’s SST reconstruction performance in masked regions. The MAE algorithm from He et al. (2022) automatically masks out random patches from the image as shown in the middle column of Fig. 2. While the extent to which these random patches represent real clouds is an open question for future study, Sect. 5 shows a comparison between masking random patches and masking with real clouds.

To extract the SST data, the full SST grid from LLC4320 is partitioned into tiles, each containing 256×256 grid points (equivalent to approximately $512 \text{ km} \times 512 \text{ km}$ in mid-latitudes). Tiles that contain land or sea ice, which are typically not observed in real-world satellite data, are excluded from the dataset. The dataset is further segregated into a training set and a validation set based on the “year” in the simulation. While neither are fed into the MAESSTRO training pipeline, the LLC4320 validation set should not be confused with the LLC2160 test set. In machine learning, validation sets are subsets of the training data used to op-

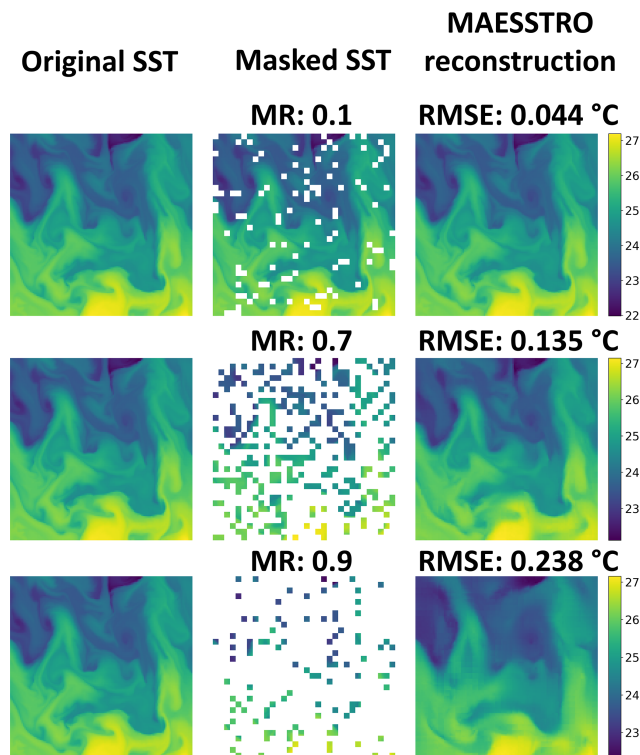


Figure 2. Example MAESSTRO result for a sample tile from the unseen validation set with masking ratios (MR; % of pixels removed) from 0.1 to 0.9. MAESSTRO is an adapted version of the MAE model that is trained on masked LLC4320 tiles.

timize hyperparameters and model architectures. Test sets, on the other hand, are completely unseen datasets that cannot be used to make any decisions regarding the model or its training process. Therefore, MAESSTRO's performance on the LLC2160 test set is a better estimate of its generalization ability on real-world unseen data.

Approximately 750 362 tiles from 2012 are allocated for training, while roughly 250 447 tiles from 2011 are assigned for validation. The use of a comprehensive and high-resolution simulation such as LLC4320 facilitates an in-depth evaluation of the MAE model's ability to capture spatial SST variability.

LLC2160 SST tiles with 128×128 tile size (approximately $512 \text{ km} \times 512 \text{ km}$ in the mid-latitudes) are used as the unseen test set. The grid spacing is twice as large as in LLC4320, resulting in a lower spatial resolution, which also tests MAESSTRO's ability to generalize across different spatial resolutions.

2.3 Data augmentation and pre-processing

The SST tiles extracted from the LLC4320 simulation outputs have dimensions of 256×256 , while the MAE model used in this paper takes input tiles with dimensions of 128×128 . To resize the tiles, a random portion of the full

tile is cropped, ranging from 20 % to 100 % of the original tile, before being resized to the final 128×128 dimensions using bicubic interpolation. This process of *data augmentation* enables the model to learn features across multiple scales while artificially increasing the number of tiles available for training.

The SST data are standardized based on statistics from the training set (2012) to have zero mean and unit variance. This is achieved by subtracting the mean (14.84°C) and dividing by the standard deviation (9.38°C). It is important to note that only SST tiles from the training set were used to calculate the statistics used for standardization to minimize any “leakage” of information from the unseen validation and test sets. Additionally, during evaluation, the data are re-scaled back to degrees Celsius.

2.4 Model training

Figure 2 shows samples of the reconstructed SST fields when testing the MAE model on an unseen SST tile from LLC4320 with different masking ratios from 10 % to 90 %. The model is able to reconstruct fine-scale SST features up to 80 %, but the sharpness of the reconstructed SST fields decreases as the masking ratio increases. In Fig. 2, random portions of the original SST tiles are masked out at a prescribed masking ratio. The two primary differences between such a masking strategy and real clouds are the following:

1. *The constant masking ratio.* Different regions of the ocean exhibit different cloud coverage; thus a model needs to “generalize” to different masking ratios.
2. *The uniform random masks.* Clouds are not distributed uniformly across the Earth and should therefore not be uniformly distributed in each input tile to the model. For example, there could be a large, continuous patch of clouds that only occupies a certain region of the tile rather than evenly occupying the entire image.

This initial investigation will focus on the model's ability to generalize to different masking ratios to address varying levels of cloud cover. The second issue of predicting under unevenly distributed clouds is the topic of a future publication.

Preliminary experiments revealed issues with generalization ability; MAESSTRO models trained on a certain masking ratio performed poorly when tested on other masking ratios, regardless of whether or not the test-time masking ratio was higher or lower. It was found that training with *random masking ratios* sampled from a uniform distribution with a mean masking ratio of 0.5 alleviated this overfitting issue. For additional regularization, MAESSTRO uses smaller ViT-tiny models, which also provides the benefit of being faster to compute than the default ViT-base MAE, which is itself already computationally efficient. Despite producing a larger model (and potentially *contributing* to overfitting), a patch

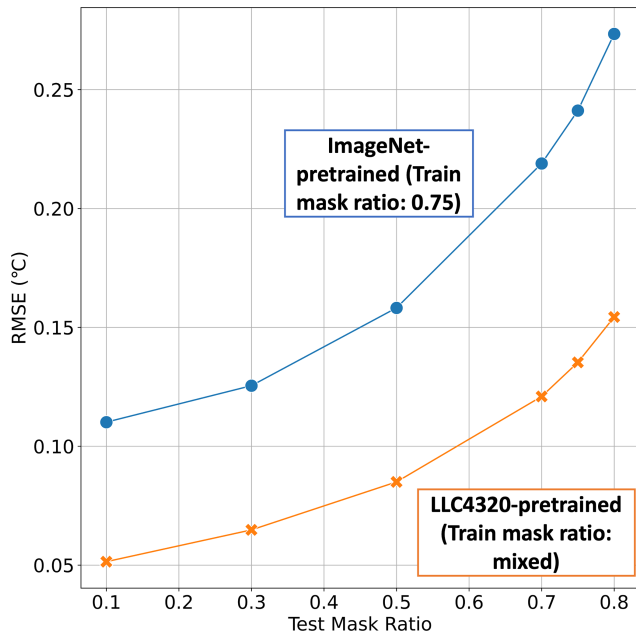


Figure 3. Comparison between MAE (ImageNet-pretrained; blue) and MAESSTRO (LLC4320-pretrained; orange) models, showing RMSE as a function of test-time masking ratio for different train-time masking ratios. RMSE is averaged across all tiles in the test set. MAESSTRO outperforms the ImageNet-pretrained MAE model from He et al. (2022). Results in this figure are shown for the LLC4320 validation set.

size of 4 for the ViT model is used to allow for finer-grained masks that can eventually be modified to capture real cloud distributions.

To establish a benchmark, we converted the single-layer SST field into a three-channel RGB image and plugged it into the existing ImageNet-pretrained MAE model in He et al. (2022). The MAE model trained on 1.2 million natural RGB images (ImageNet) with a 75 % masking ratio by He et al. (2022) did not exhibit this overfitting issue. Figure 3 shows that the MAESSTRO model trained on ocean simulation LLC4320 consistently outperforms the ImageNet-pretrained model at all test-time masking ratios evaluated. This model uses a ViT-tiny with a patch size of 4 and is trained with mixed/random masking ratios.

2.5 Evaluation metrics

We use three metrics to assess MAESSTRO's performance: root mean square error (RMSE), spatial correlation, and spatial coherence. RMSE values are only calculated for reconstructed pixels, i.e., pixels that have been randomly masked out as described in Sect. 2.4.

The second metric, spatial correlation, is calculated using Pearson's correlation coefficient for flattened the ground truth and predicted sea surface temperature (SST) fields. A smaller RMSE value typically results in a higher correlation,

Table 1. Delineation of wavenumber k into two distinct bands to capture large-scale and small-scale features. Wavenumber is inversely related to spatial scale; smaller k values represent larger length scales and vice versa.

Scale	Wavenumber, k (cycles per pixel)	Pixel scale (no. pix- els)	Length scale in LLC2160 (km)
Small	0.1–0.2	5–10	20–40
Large	0.01–0.06	17–100	70–400

but this is not always true across all spatial scales and structures.

The third metric, spatial coherence, evaluates reconstruction performance at different spatial scales. Coherence between true and reconstructed values is defined as

$$C(k) = \frac{|CSD(T, \hat{T})|^2}{\sigma^2(T)\sigma^2(\hat{T})}, \quad (1)$$

where T and $\hat{T}$ are the ground truth and reconstructed SST values, respectively, $CSD(T, \hat{T})$ is the cross-spectral density along the x axis (each row), and $\sigma^2(T)$ is the variance of T along the same x axis (128 pixels). Note that an evaluation along the y axis provides the same statistical results considering the large ensemble of images used here. For each pair of T and $\hat{T}$, $C(k)$ is the average of the coherence vector over all 128 rows, with k ranging between 0 and $1/2$ (cycles per pixel). $C(k)$ is linked to spatial scales by definition and is a vector function of the wavenumber, k . We delineate k into small scales and large scales, as shown in Table 1.

To evaluate MAESSTRO's performance on SST fronts (which are related to spatial temperature gradients), we compute each metric on the SST spatial gradient field (∇T , $\nabla \hat{T}$) in addition to the SST field (T , $\hat{T}$). Here, the SST gradient is defined as the L2 norm of temperature gradients in the x and y directions.

$$\|\nabla T\|_2 = \sqrt{T_x^2 + T_y^2} \quad (2)$$

3 Evaluation on a sample SST tile from LLC2160

We compare MAESSTRO's performance on the LLC2160 test set with the Kriging method Matheron (1963), a commonly employed gridding technique for irregular geospatial analysis and the foundation of a widely used SST product (Reynolds et al., 2007), as well as cubic radial basis function (cubic RBF) interpolation, which is less popular but surprisingly outperformed Kriging methods in our case (details are in Sect. 3.2 and Table 2). Two variogram models were used in Kriging method: linear and Gaussian. The linear variogram

Table 2. Evaluation metrics for the single-tile example shown in Figs. 4 and 5 for different reconstruction methods. The RMSE was calculated as the spatial standard deviation after removing the linear trends in the zonal and meridional directions. The correlation is the simple Pearson's correlation coefficient of the flattened (2D fields to 1D vector) anomaly fields. Our results are highlighted in bold.

Method	RMSE (°C)	Correlation	RMSE gradient (°C per pixel)	Correlation gradient
MAESSTRO (ours)	0.0229	0.990	0.00950	0.921
Cubic RBF	0.0342	0.978	0.0127	0.856
Linear-Kriging	0.0549	0.938	0.0186	0.606
Gaussian-Kriging	0.0562	0.935	0.0184	0.633

does not have a range, but the 128×128 image size effectively limits the range of the variogram at 128 pixels. The linear variogram is dynamically calculated based on each provided masked SST field. The Gaussian variogram has 0.1°C , 20 pixels, and 0.001°C for the sill, range, and nugget parameters, respectively. The cubic RBF fits a cubic function for each data point with a basis function $\phi(r) = r^3$, where r represents the distance of a center point to a point with values. The cubic RBF exhibits a smooth surface/feature that is suitable for the turbulence nature shown by the SST images. The LLC2160 SST tiles consist of 128×128 grid points (approximately $512\text{ km} \times 512\text{ km}$ in the mid-latitudes). The grid spacing is twice as large as that in LLC4320, thereby allowing us to test MAESSTRO's ability to generalize across different spatial resolutions.

Given an image with a total of M pixels, where N pixels are missing, the missing pixels are filled using cubic-RBF interpolation. The value at each missing pixel x_j is estimated by

$$T(x_j) = \sum_{i=1}^{M-N} c_i \phi(\|x_j - x_i\|) \text{ for } j = 1, \dots, N,$$

where c_i denotes the coefficients associated with the known pixels x_i , and $\phi(r) = r^3$ is the cubic radial basis function. These coefficients c_i are determined through an optimization process based on the known pixel values.

To introduce the evaluation metrics and the algorithms used in this paper, we first analyze reconstructions on an example SST tile where 80 % of the original values were removed at random. The masked and original ground truth SSTs are shown in Fig. 4a and b.

3.1 Qualitative evaluation

In this example, the MAESSTRO reconstruction accurately captures the primary strong SST front, as well as smaller-scale structures such as the two filaments in the bottom-left corner ($j = 0; i = 50$). MAESSTRO preserves the anisotropic frontal structures without over-smoothing in the cross-frontal direction, as seen in Fig. 4c. In contrast, direct interpolations using cubic-RBF interpolation or Kriging with linear and Gaussian variogram models fail to capture the primary sharp front, resulting in smoothed and blurry reconstructions

with spurious locally generated “bull’s eye” features, such as the one at $j = 50, i = 50$, which are typical of objective analysis.

The spatial gradient of SST, $\|\nabla T\|_2$ (as opposed to SST itself), effectively highlights the small-scale frontal structures that are the focal point of this study. Qualitative comparisons of the reconstructed SST gradients show that MAESSTRO reconstructs the sharp frontal structure more accurately than cubic-RBF interpolation, as shown in Fig. 5.

3.2 Quantitative analysis

The quantitative results shown in Table 2 are consistent with the qualitative assessment in Figs. 4 and 5. The MAESSTRO reconstruction exhibits a lower RMSE and higher correlation than the other methods. However, the performance difference between MAESSTRO and the other methods is more pronounced for RMSE than for correlation. This is understandable because correlation primarily captures large-scale features that can be retrieved reasonably well using conventional methods. MAESSTRO's advantage over conventional methods lies in its ability to reconstruct smaller-scale features, which is better quantified by the correlation in the SST *gradient* field, denoted as $\|\nabla T\|_2$. Specifically, the correlation of the SST gradient of MAESSTRO's reconstruction with the ground truth is 0.92, whereas the interpolation methods yield correlation values from 0.63 to 0.85. Cubic-RBF interpolation is the best-performing conventional method and will serve as a benchmark in the global evaluation discussed in subsequent sections.

The square of the spatial coherence, $C(k)$, elucidates each algorithm's performance at various spatial scales. In this example tile, the MAESSTRO reconstruction exhibits significantly higher coherence for spatial scales larger than around 5 pixels. This indicates that MAESSTRO accurately reproduced the frontal structures in Fig. 4 down to a 5-pixel scale. On the other hand, the SST reconstruction from cubic-RBF interpolation has the highest coherence at scales smaller than around 3 pixels. The comparison of $C(k)$ across different algorithms is shown in Fig. 6a.

Furthermore, the power spectral density (PSD) calculated from the MAESSTRO reconstruction matches the true PSD down to a wavelength of about 4 pixels (wavenumber 0.25 cycles per pixel). Below 4 pixels, the reconstructed SST

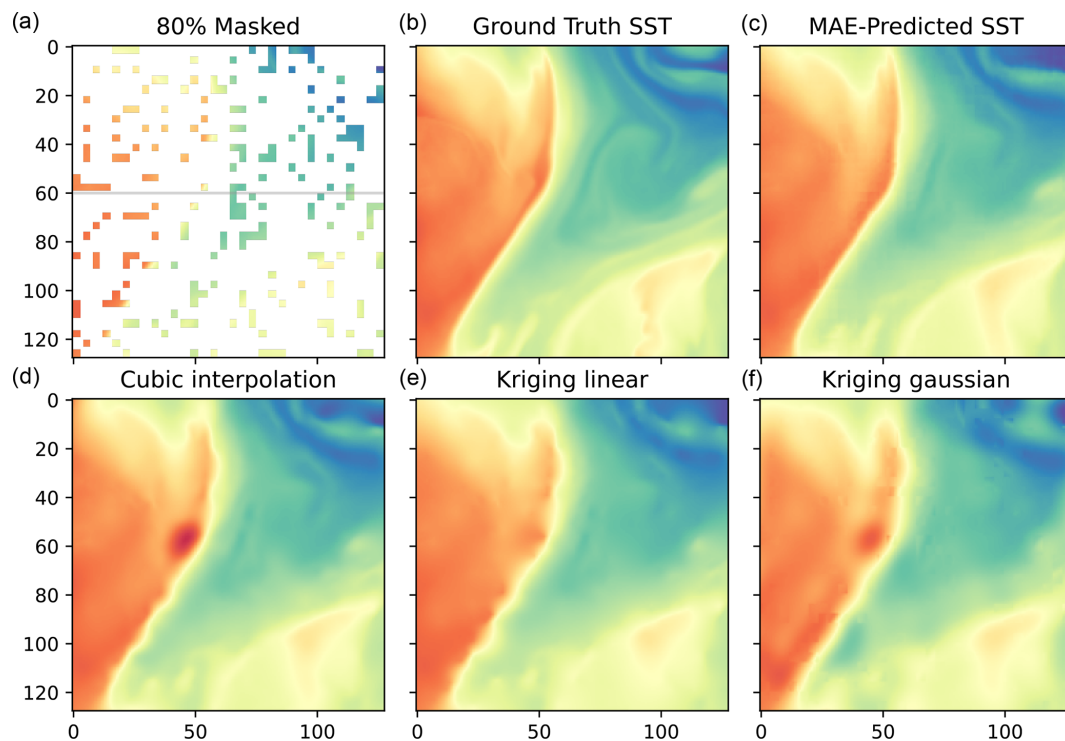


Figure 4. Example of a masked SST tile from the unseen LLC2160 test set with 80 % missing/masked values (a), the ground truth from the model (b), MAE-reconstructed SST based on the masked field (c), and the reconstructed SST using three interpolation schemes (d–f), from left to right: cubic-RBF interpolation, Kriging interpolation with a linear kernel, and Kriging interpolation with a Gaussian kernel.

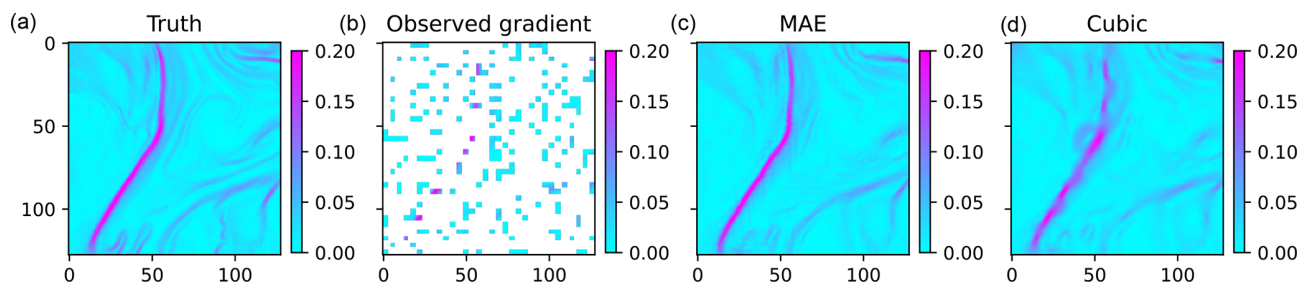


Figure 5. The $\|\nabla T\|_2$ associated with the SST field in Fig. 4 with a unit of $^{\circ}\text{C}$ per pixel. The panels from left to right present (a) the true SST gradient, (b) observed gradient from the masked SST field, (c) SST gradient from MAE reconstruction, and (d) SST gradient from cubic-RBF interpolation, respectively.

exhibits a higher PSD than the ground truth, indicating the presence of excessive noise in the small-scale reconstruction. The PSD values are shown in Fig. 6b. This sudden drop in performance at the 4-pixel scale can be attributed to the patch size utilized by MAESSTRO, as the ViT-tiny model processes the image based on 4×4 patches. The signal-to-noise ratio within each 4×4 patch might be slightly higher compared to the overall image. Conversely, the cubic-RBF interpolation reconstruction exhibits a lower PSD than the ground truth below 4 pixels, which may explain its higher coherence at those scales.

To further explore the issue of noise at small scales, the high-pass-filtered SST along $j = 60$ of the example tile is

shown in Fig. 6d. The small-scale SST values at that latitude exhibit higher grid noise in the MAESSTRO reconstruction than the other interpolation methods. However, it is important to note that the SST amplitude is extremely small ($\sim 0.01^{\circ}\text{C}$) at the 4-pixel level, in contrast to the 1°C amplitude across the major SST front (Fig. 6c) spanning approximately 20 pixels ($x = [40 \dots 60]$).

4 Global validation results on LLC2160

In the preceding section, we presented each evaluation metric and linked these numerical values with corresponding vi-

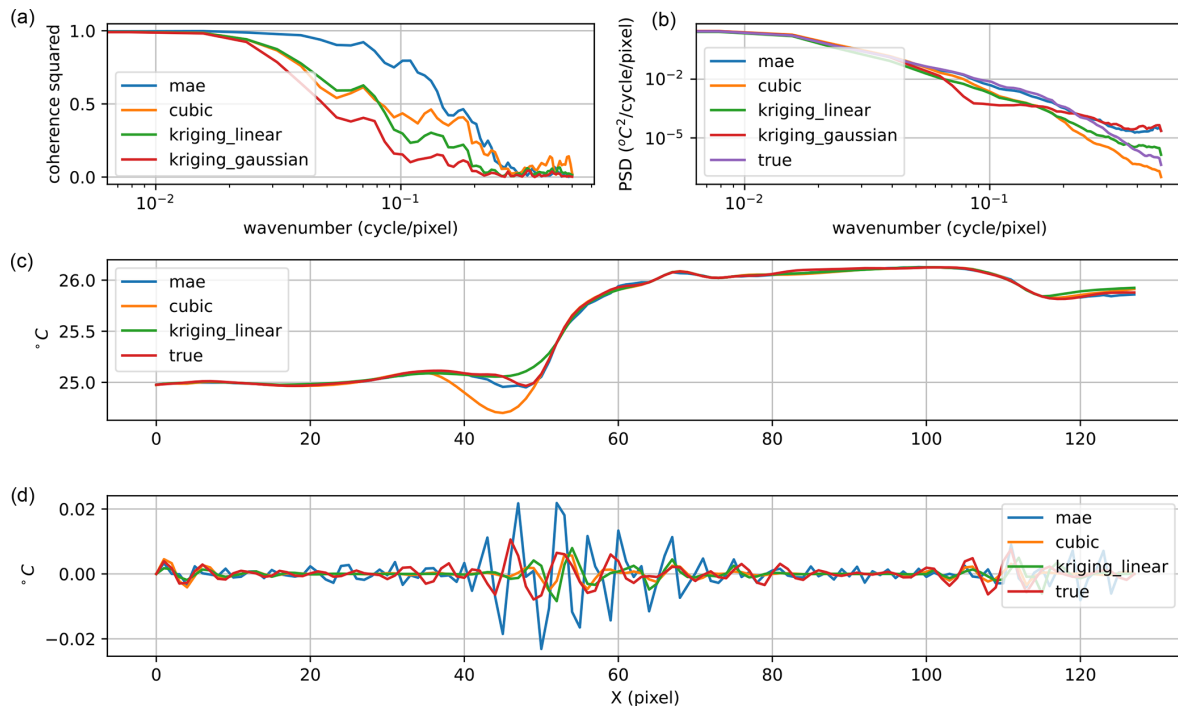


Figure 6. (a) Coherence between ground truth and reconstructed SST. (b) The power spectral density of the ground truth (purple) and reconstructed SST fields. (c) The SST profile along $j = 60$ marked by the gray line in Fig. 4. (d) The high-pass-filtered SST with a cutoff wavelength threshold at 6 pixels. Note that the variability of the small scale (d) is less than 2 % of the total SST variability (c).

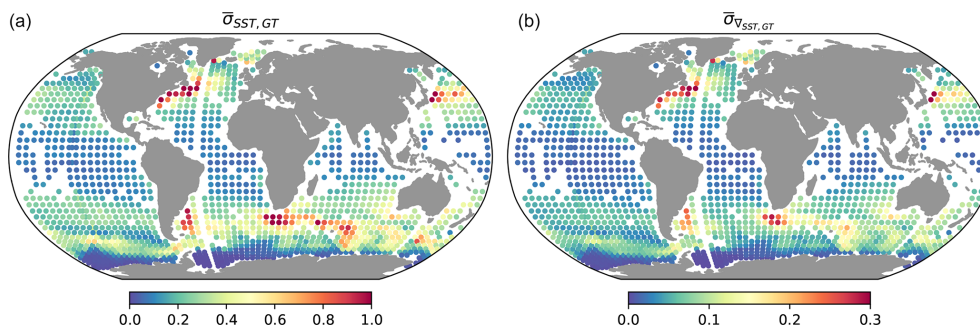


Figure 7. Global maps showing time-averaged standard deviation within a $128 \text{ pixel} \times 128 \text{ pixel}$ domain for SST (a) and the SST spatial gradient (b; defined as $\|\nabla \text{SST}\|_2$), derived from 83 global SST snapshots from the LLC2160 dataset. These values serve as a benchmark for reconstructed SST maps using MAESSTRO and other methods.

sual representations. In this section, we assess MAESSTRO's global performance by presenting the evaluation metric values on a global map and comparing them with cubic-RBF interpolation results as a benchmark. We utilized 83 global SST snapshots from the LLC2160 dataset with a 10 d interval between snapshots, spanning a period of over 2 years. We calculated the statistics for each individual snapshot and subsequently averaged them over the 2-year period.

To evaluate MAESSTRO's ability to accurately reconstruct the *global SST variability*, we use the standard deviation of SST within each 128×128 tile as a measure of SST variability. Prior to computing the standard deviation,

we remove zonal and meridional linear trends to eliminate large-scale signals that can be accurately captured by cloud-immune microwave satellite sensors, thereby allowing us to concentrate on smaller spatial scales.

The global geographic distribution of ground truth SST and SST gradient ($\|\nabla T\|_2$) shows distinct but typical patterns that are associated with strong ocean currents, as we can see in Fig. 7. The regions with large amplitude values in both fields are often found in areas where these currents occur, such as the western boundary currents in the northwestern Pacific (Kuroshio), northwestern Atlantic (Gulf Stream), southwestern Atlantic (Malvinas), and the Agulhas retroflec-

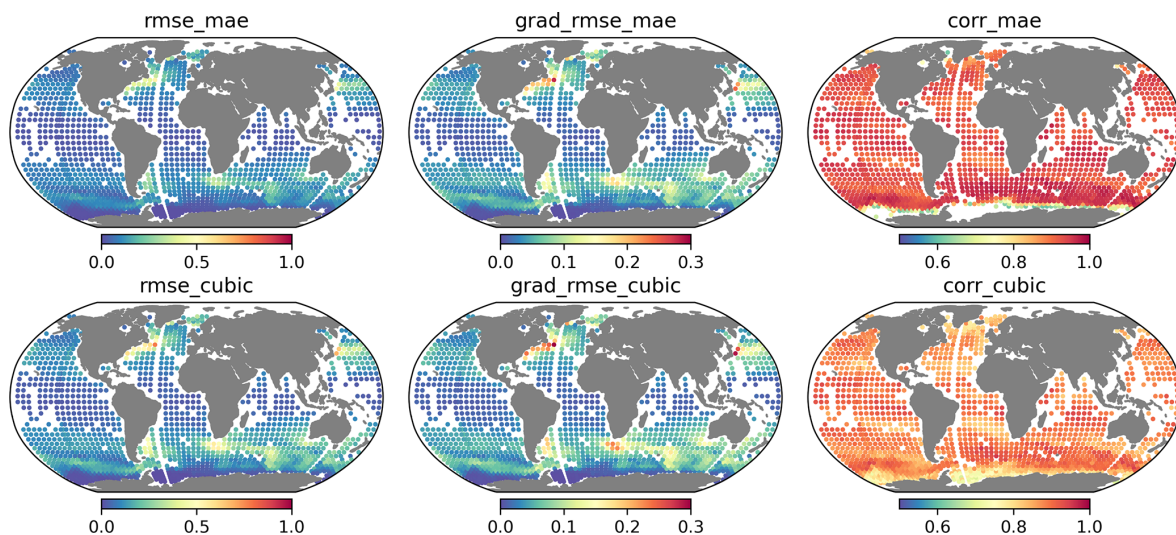


Figure 8. Global maps showing the *reconstructed* standard deviation within a 128 pixel $\times$ 128 pixel domain for SST (left), the SST spatial gradient ($\|\nabla \text{SST}\|_2$) (middle), and the correlation coefficients (right).

tion, as well as the Antarctic Circumpolar Currents downstream of major topographic features such as the Kerguelen Plateau. These large-amplitude regions are shown as red dots in Fig. 7.

In these regions, the standard deviation of SST ranges from 0.6–1.0 °C, while the SST gradient typically has a standard deviation of 0.1–0.3 °C per pixel calculated over an area of 128 $\times$ 128 pixels (approximately 512 km $\times$ 512 km in mid-latitudes). The Antarctic region (blue region around the South Pole in Fig. 8) exhibits very low SST variability due to sea ice cover. As mentioned in Sect. 2.2, tiles under sea ice were removed from the LLC4320 training set. These values serve as benchmarks against which each method's ability to reconstruct global SST variability will be compared.

Figure 8 shows the RMSE and spatial correlation for MAESSTRO (top row) and the cubic-RBF interpolation method (bottom row). Consistent with the findings in Sect. 3.2, the MAESSTRO reconstruction exhibits smaller RMSEs for both the SST and SST gradient, along with higher correlations compared to the cubic-RBF interpolation. The relatively smaller amplitude of the RMSE compared to the ground truth (as depicted in Fig. 7 using the same color scale) indicates that the reconstruction is meaningful. That is, the error magnitude of reconstructed SST is smaller than the actual SST signal. The higher correlation values further validate the accuracy of the reconstruction.

Both methods show high spatial correlation (> 0.7). However, in the case of MAE, an exception arises in the Antarctic region where no training images were available for under-ice conditions. The temperature images used in this evaluation were not sea surface temperature but *under-ice temperature*. While cubic-RBF interpolation can fill in the missing values, MAESSTRO encounters out-of-domain (OOD) challenges

due to the absence of training images containing under-sea ice.

When comparing the global distributions of RMSE and correlation across MAESSTRO and all baseline methods, MAESSTRO's reconstructions exhibit the lowest mean RMSE and the highest correlation, as illustrated in the letter-value plot in Fig. 9. Contrary to what is shown in Table 2, there is a wider gap in correlation performance between MAESSTRO and cubic-RBF interpolation (the best-performing baseline) compared to RMSE. More importantly, MAESSTRO still exhibits a larger advantage in the correlation between the reconstructed and ground truth SST gradient, $\|\nabla T\|_2$, thus highlighting its ability to accurately reconstruct smaller-scale features across the global ocean.

To further differentiate between small and large scales, we now investigate the global maps of coherence, which have been averaged across two wavelength bands, specifically, 5–10 pixels (small scales) and 17–100 pixels (large scales), as given in Table 1. Both SST and SST gradient statistics are presented.

Figure 10 shows the squared coherence evaluation for MAE (Fig. 10a–d) and cubic-RBF interpolation (Fig. 10e–h). Although nearly all coherence values are significant at a 0.01 significance level, MAE consistently yields better performance globally. Generally, coherence is smaller at small scales for the same quantity (SST or SST gradient) because capturing small-scale features is more challenging. At large scales, SST reconstructions (the first column) exhibits greater coherence with the truth than the SST gradient (the third column). However, at small scales, the MAESSTRO prediction of the SST gradient (rightmost column) demonstrates higher coherence than the prediction of SST itself (the second col-

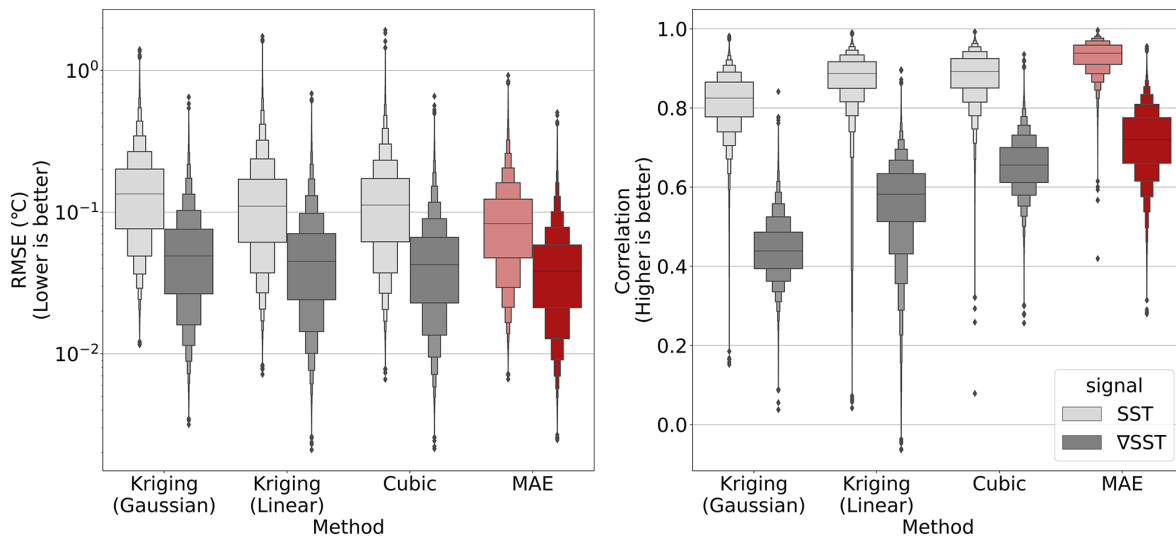


Figure 9. Distributions of RMSE and correlation metrics across predictions for all tiles in the SST2160 test set. MAE results are highlighted in red. Note that RMSE results are shown on a logarithmic scale. MAE's default random patch masking scheme (i.e., not real clouds) from He et al. (2022) was used to obtain these results.

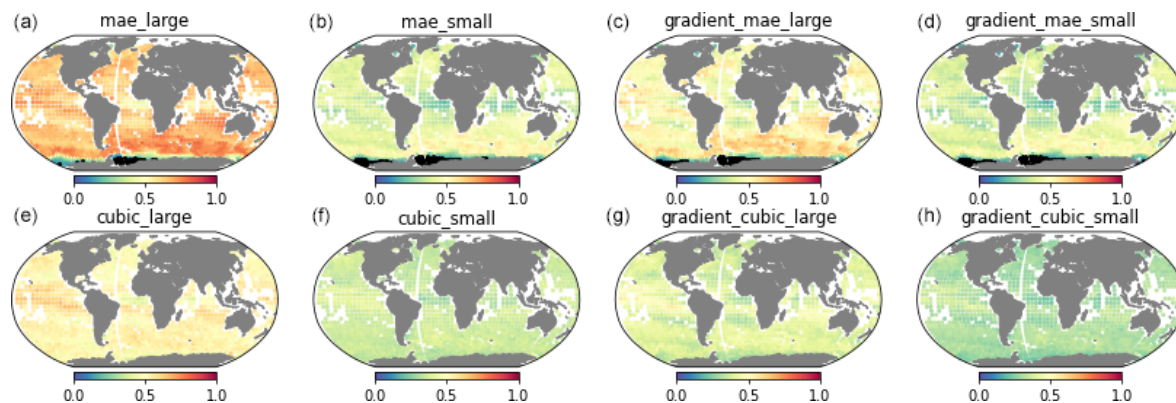


Figure 10. The squared coherence between reconstruction and ground truth, $C(k)$, for MAESSTRO (a–d) and cubic-RBF interpolation (e–h). The first (a, e) and third (c, g) columns show the results for the large scale (17–100 pixels), while the second (b, f) and fourth (d, h) columns show the results for the small scale (5–10 pixels). The coherence squared is significant at 0.1 using a 0.01 significant level with 83 independent samples in the time average. The insignificant values are blacked out, mostly near the Antarctic.

umn). This emphasizes the efficacy of MAE in retrieving small-scale SST fronts.

5 Evaluation on a satellite SST tile

Our primary objective is to demonstrate the practicality of using MAE to reconstruct ocean fronts. As a first step, we employed synthetic SST fields generated from a high-resolution $1/48^\circ$ numerical ocean simulation for both training and validation. This allowed us to establish a controlled environment to evaluate the effectiveness of our approach. To substantiate our methodology, we tested it using real satellite sea surface temperature (SST) data from the Suomi NPP Visible Infrared Imaging Radiometer Suite (VIIRS) in the California Current

region, specifically at coordinates 35°N , 125°W on 16 January 2021 (Fig. 11a–c). This 200 km by 200 km SST tile was extracted from the high-resolution (750 m) VIIRS SST level-2P product (JPL/OBPG/RSMAS, 2020) and regridded to 128×128 pixels.

We examined masking ratios of 0.1, 0.3, 0.5, 0.7, and 0.9 using 450 unique random masks for each ratio. Across these masking ratios, MAESSTRO produces RMSE trends consistent with those in Fig. 3, achieving an RMSE of less than 0.1°C for masking ratios of up to 0.9 and 0.05° for masking ratio of 0.1, as seen in Fig. 12.

The coherence and power spectral density further evaluate the fidelity of MAESSTRO's reconstructions across various spatial scales. Two distinct masking ratios, namely 0.3 and

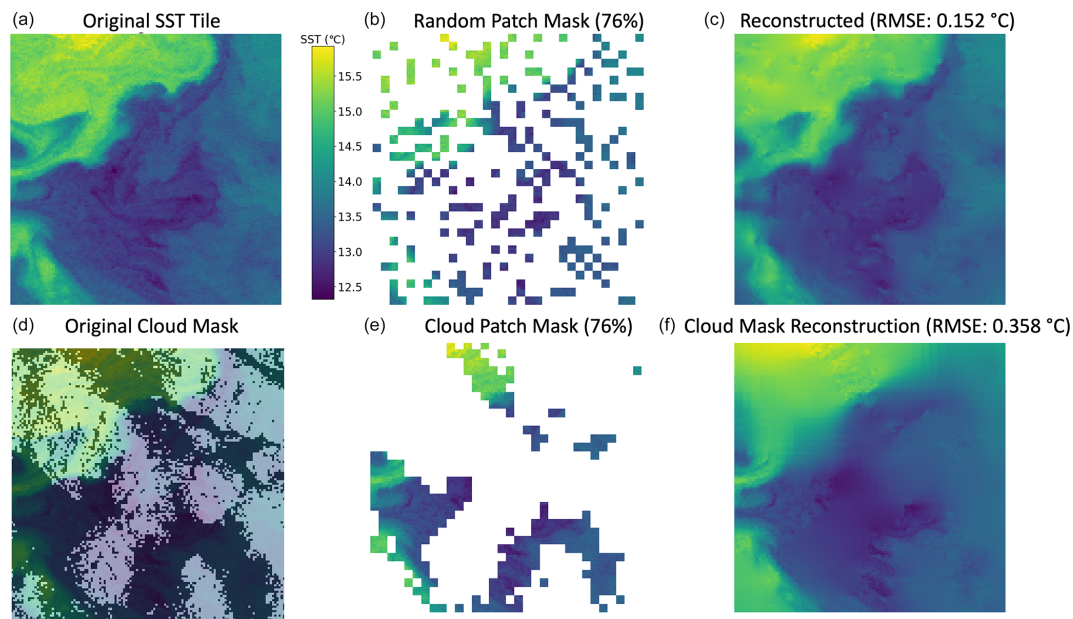


Figure 11. (a) The SST over a $200\text{ km} \times 200\text{ km}$ area in the California Current region from the VIIRS Suomi NPP Level-2P product (JPL/OBPG/RSMAS, 2020). The slanted horizontal strips are satellite artifacts. (b) Masked SST tile using MAE's default random patch masking technique. (c) MAESSTRO reconstruction from the randomly masked middle panel with a masking ratio of 0.76. (d) VIIRS/SNPP cloud mask overlaid over the VIIRS SST tile. (e) Masked SST tile using the VIIRS/SNPP cloud mask. (f) MAESSTRO reconstruction from the realistic cloud-masked middle panel with a masking ratio of 0.76.

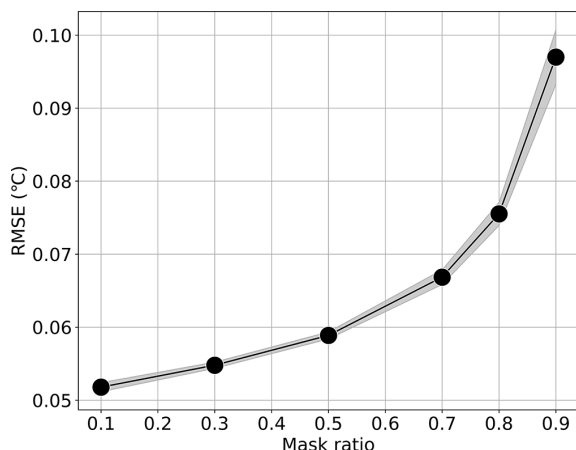


Figure 12. Mean RMSE with 95 % confidence interval for LLC4320-pretrained ViT-tiny patch 4 model evaluated at 450 random masks on the $512\text{ km} \times 512\text{ km}$ VIIRS tile for each test-time masking ratio.

0.8, are shown as an example in Fig. 13. For instance, at a masking ratio of 0.3, the reconstructed data (blue) closely align with the actual VIIRS SST in terms of the spectral energy (Fig. 13b) while maintaining a high level of coherence (Fig. 13a). At a masking ratio of 0.8, where gaps between available data points are substantial, the reconstruction begins to lose precision at small scales (wavelengths less than approximately 10 km), a trend evident in both measures,

but maintains high accuracy for large scales ($> 10\text{ km}$ wavelengths).

MAESSTRO's performance on VIIRS data with a real cloud mask is shown in Fig. 11d–f. We selected a cloud mask on 24 March 2012 from the same L2P dataset such that the masking ratio matched the random patch mask in Fig. 11a–c. Perhaps unsurprisingly, MAESSTRO produces higher RMSE when the SST tile is occluded with a realistic cloud mask compared to the random patch mask used by the MAE algorithm. While both masks have the same masking ratio, the cloud mask has larger contiguous regions with missing data. This indicates the need to train MAESSTRO on SST data with real cloud masks, which will be the focus of future work.

In summary, MAESSTRO demonstrates a globally consistent performance on simulated data at two resolutions ($\sim 2\text{ km}$ and $\sim 4\text{ km}$ grids) and on the satellite SST data at a pixel resolution of 750 m. These evaluations confirm the method's effectiveness and its insensitivity to variations in spatial resolution and scales, echoing the concept of dynamic similarity found in fluid dynamics.

6 Discussion

In our study, we employed a large set of tiles from a numerical simulation for training, but it is worth highlighting that adding more real-world data could likely improve the performance of the MAE model. We generated synthetic cloud

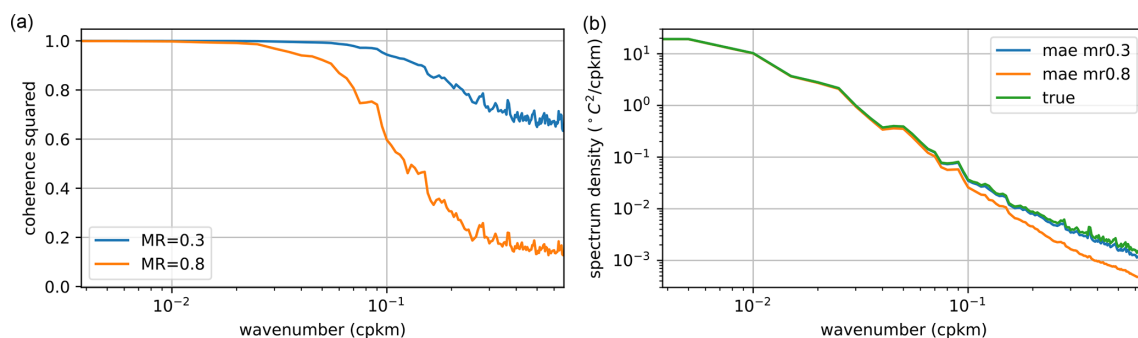


Figure 13. (a) The coherence between the VIIRS SST and the MAESSTRO reconstruction for masking ratio 0.3 (blue) and 0.8 (orange). (b) The power spectral density of the VIIRS SST (green) and the reconstruction for the same masking ratios.

cover randomly, but actual cloud cover can display more diverse patterns. Moreover, we have not thoroughly examined the effect of random noise on MAE predictions, even though such noise often appears in level-2 SST products. Our test on VIIRS SST provides preliminary evidence of MAESSTRO's capacity to handle noisy images. Here, we have evaluated one case with real cloud cover, but future studies should consider realistic cloud cover with more robust statistics and address random noise to better represent satellite SST images.

The artificially generated cloud mask employed in this study resembles spotted cloud patterns, such as those produced by altocumulus clouds with relatively uniform spatial distribution. Nonetheless, it is important to acknowledge that significant errors are anticipated in areas with continuous cloud masks over an extensive area, as shown in Fig. 11. This is due to the model's inability to establish connections between data points across large spatial distances. Small-scale features may be lost if the gap created by clouds is too large. Future investigations could explore methods for integrating complementary information, such as incorporating low-resolution, cloud-free microwave SST data and/or cloud-free sea surface height data from altimeters.

Another key aspect of our study is the implementation of scaling augmentation during model training. As mentioned in Sect. 2.4, at every iteration, the model is trained on different crops from the original source tile. This enables the model to be robust to different scales and relatively proficient in reconstructing structures at various spatial scales. Consequently, MAESSTRO possesses the potential for application in other geophysical fields such as ocean color remote sensing images, which also display complex structures generated by the same flow fields. The MAE model's adaptability in managing different spatial scales makes it a promising instrument for retrieving small-scale ocean processes.

Lastly, MAESSTRO notably outperforms conventional methods in efficiency, being over 5000 times faster. For instance, based on a single CPU, MAESSTRO can complete a global reconstruction with 1382 tiles of 128×128 size within 37 s, while the same calculation using cubic-RBF interpolation and Kriging method took approximately 24 h.

This combined benefit of increased speed and enhanced accuracy becomes particularly vital in an era dominated by large datasets.

7 Conclusions

Small-scale SST fronts are integral to marine physics and have a considerable influence on Earth's climate system. State-of-the-art infrared satellite SSTs can resolve these fronts at a resolution as fine as 750 m grid. Nonetheless, due to the inability of infrared to penetrate clouds, high-resolution SST observations are not consistently available across all weather conditions. The retrieval of obscured, high-resolution SST data has posed a persistent challenge for satellite oceanography, as existing techniques have struggled to make efficient predictions without compromising the original high-resolution information.

In this paper, we demonstrate that employing a ViT model in conjunction with a masked autoencoder significantly outperforms traditional optimal interpolation approaches. While there remains scope for enhancement, our results indicate that machine learning models, predominantly used in computer vision for visual representation, can also be adapted for quantitative purposes in satellite oceanography. We leveraged numerical simulations for training and evaluation. As a next step, we plan to extend our methodology through further training and evaluation using high-resolution SST datasets from the VIIRS and MODIS sensors as well as other supporting satellite measurements including lower-resolution microwave SST and sea surface height from conventional altimeters and the just-launched Surface Water and Ocean Topography (SWOT) mission.

Appendix A: Hyperparameters used to train MAESSTRO

Table A1. MAESSTRO model configuration/parameters. We adapted the masked autoencoders (MAE) GitHub repository from (He et al., 2022) to operate on single-channel SST tiles. Compared to the original MAE, MAESSTRO is 10 % the size and takes inputs of $1 \times 128 \times 128$ rather than $3 \times 224 \times 224$, where the first value corresponds to the number of channels, and the second and third values refer to the input height and width, respectively. The smaller network size contributes to MAESSTRO’s computational efficiency. MAESSTRO is trained on randomly masked synthetic SST tiles from the LLC4320 simulation.

Parameter	Value
ViT variant	ViT-tiny
Patch size	4×4
Encoder embedding dimension	192
Encoder depth	12
Encoder number of attention heads	3
Patch size	4×4
Decoder embedding dimension	192
Decoder depth	12
Decoder number of attention heads	3
Input size	128×128
Number of parameters	11 114 896
Masking ratio	Randomly distributed between 0.1 and 0.9
Learning rate	0.00015
Batch size	1968 (82 per GPU)
Number of training tiles	750 362
Number of validation tiles	250 447
Number of epochs	300
Number of GPUs	24 NVIDIA A100 GPUs with 40 GB RAM each
Training time	20 h, 21 min, 32 s

Code availability. Code has been made available at this DOI: <https://doi.org/10.5281/zenodo.13799966> (Chen and Goh, 2024).

Data availability. The global MITgcm simulations used in this study, llc4320 and llc2160, are available at https://data.nas.nasa.gov/ecco/eccodata/llc_4320/ (NASA, 2023a) and https://data.nas.nasa.gov/ecco/eccodata/llc_2160/ (NASA, 2023b), respectively.

Author contributions. EG implemented the MAE solution, developed the MAESSTRO code, and trained/evaluated MAESSTRO. AY carried out a part of the model validation. JW conceptualized the problem, conducted the global llc2160 and VIIRS evaluation, and drafted the manuscript. BW is the principal investigator of the project and ensured the rigor of the study. All authors contributed to interpretation of the results and writing and finalization of the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Special issue statement. This article is part of the special issue “Special Issue for the 54th International Liège Colloquium on Machine Learning and Data Analysis in Oceanography”. It is not associated with a conference.

Acknowledgements. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). The High Performance Computing resources used in this investigation were provided by funding from the JPL Information and Technology Solutions Directorate. We acknowledge the independent but similar research effort conducted by Agabin et al. (2024). We thank Dimitris Menemenlis and the MITgcm team for creating llc2160 and llc4320 and making it available. We thank Christopher Henze and his team at NASA Ames for assisting in data

access. LLC2160 and LLC4320 were run on the Pleiades cluster at NASA Ames.

Financial support. This research has been supported by the National Aeronautics and Space Administration (grant no. 80NM0018D0004) under the Advanced Information Systems Technology (AIST) program (funding no. NNH21ZDA001N-AIST). Jinbo Wang was partially funded by the NASA Physical Oceanography Program and the SWOT mission.

Review statement. This paper was edited by Alexander Barth and reviewed by two anonymous referees.

References

- Agabin, A., Prochaska, J. X., Cornillon, P. C., and Buckingham, C. E.: Mitigating masked pixels in a climate-critical ocean dataset, *Remote Sens.*, 16, 2439, <https://doi.org/10.3390/rs16132439>, 2024.
- Alvera-Azcárate, A., Barth, A., Sirjacobs, D., Lenartz, F., and Beckers, J.-M.: Data Interpolating Empirical Orthogonal Functions (DINEOF): a tool for geophysical data analyses, *Mediterr. Mar. Sci.*, 12, 5–11, <https://doi.org/10.12681/mms.64>, 2011.
- Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M.: DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations, *Geosci. Model Dev.*, 13, 1609–1622, <https://doi.org/10.5194/gmd-13-1609-2020>, 2020.
- Barth, A., Alvera-Azcárate, A., Troupin, C., and Beckers, J.-M.: DINCAE 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations, *Geosci. Model Dev.*, 15, 2183–2196, <https://doi.org/10.5194/gmd-15-2183-2022>, 2022.
- Belkin, I. M.: Remote Sensing of Ocean Fronts in Marine Ecology and Fisheries, *Remote Sens.*, 13, 883, <https://doi.org/10.3390/rs13050883>, 2021.
- Čatipović, L., Matić, F., and Kalinić, H.: Reconstruction Methods in Oceanographic Satellite Data Observation – A Survey, *Journal of Marine Science and Engineering*, 11, 340, <https://doi.org/10.3390/jmse11020340>, 2023.
- Chen, X. and Goh, E.: jpl-slice/maestro: Initial release for Ocean Sciences MAESSTRO paper (v0.1.0), Zenodo [code], <https://doi.org/10.5281/zenodo.13799966>, 2024.
- Chin, T. M., Vazquez-Cuervo, J., and Armstrong, E. M.: A multi-scale high-resolution analysis of global sea surface temperature, *Remote Sens. Environ.*, 200, 154–169, 2017.
- Dong, J., Fox-Kemper, B., Zhang, H., and Dong, C.: The Scale and Activity of Symmetric Instability Estimated from a Global Submesoscale-Permitting Ocean Model, *J. Phys. Oceanogr.*, 51, 1655–1670, 2021.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houshy, N.: An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale, arXiv [preprint], arXiv:2010.11929, 2020.
- Fablet, R., Viet, P. H., and Lguensat, R.: Data-driven Models for the Spatio-Temporal Interpolation of satellite-derived SST Fields, *IEEE Transactions on Computational Imaging*, 3, 647–657, 2017.
- Fox-Kemper, B., Ferrari, R., and Hallberg, R.: Parameterization of Mixed Layer Eddies. Part I: Theory and Diagnosis, *J. Phys. Oceanogr.*, 38, 1145–1165, 2008.
- Goh, E., Chen, J., and Wilson, B.: Mars Terrain Segmentation with Less Labels, in: 2022 IEEE Aerospace Conference (AERO), 5–12 March 2022, Yellowstone Conference Center, Big Sky, Montana, USA, IEEE, 1–10, 2022.
- Han, Z., He, Y., Liu, G., and Perrie, W.: Application of DINCAE to Reconstruct the Gaps in Chlorophyll-*a* Satellite Observations in the South China Sea and West Philippine Sea, *Remote Sens.*, 12, 480, <https://doi.org/10.3390/rs12030480>, 2020.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R.: Masked Autoencoders Are Scalable Vision Learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18–24 June 2022, New Orleans, LA, USA, 16000–16009, 2022.
- JPL/OBPG/RSMAS: GHRSST Level 2P Global Sea Surface Skin Temperature from the Visible and Infrared Imager/Radiometer Suite (VIIRS) on the Suomi-NPP satellite (GDS2), Ver. 2016.2., PO.DAAC, CA, USA [data set], last access: 1 May 2022, <https://doi.org/10.5067/GHRSST-2PJ62>, 2020.
- Jung, S., Yoo, C., and Im, J.: High-Resolution Seamless Daily Sea Surface Temperature Based on Satellite Data Fusion and Machine Learning over Kuroshio Extension, *Remote Sens.*, 14, 575, <https://doi.org/10.3390/rs14030575>, 2022.
- Krasnopolsky, V., Nadiga, S., Mehra, A., Bayler, E., and Behringer, D.: Neural networks technique for filling gaps in satellite measurements: Application to ocean color observations, *Comput. Intel. Neurosci.*, 2016, 6156513, <https://doi.org/10.1155/2016/6156513>, 2016.
- Lévy, M., Franks, P. J., and Smith, K. S.: The role of submesoscale currents in structuring marine ecosystems, *Nat. Commun.*, 9, 4758, <https://doi.org/10.1038/s41467-018-07059-3>, 2018.
- Mahadevan, A.: The Impact of Submesoscale Physics on Primary Productivity of Plankton, *Annu. Rev. Mar. Sci.*, 8, 161–184, 2016.
- Mahadevan, A. and Tandon, A.: An analysis of mechanisms for submesoscale vertical motion at ocean fronts, *Ocean Modell.*, 14, 241–256, 2006.
- Matheron, G.: Principles of geostatistics, *Econ. Geol.*, 58, 1246–1266, <https://doi.org/10.2113/gsecongeo.58.8.1246>, 1963.
- McWilliams, J. C.: Submesoscale currents in the ocean, *P. Roy. Soc. A-Math. Phys.*, 472, 20160117, <https://doi.org/10.1098/rspa.2016.0117>, 2016.
- NASA: Global MITgcm simulation – llc4320, Ecco Data Portal [data set], https://data.nas.nasa.gov/ecco/eccodata/llc_4320/, last access: 1 March 2023a.
- NASA: Global MITgcm simulation – llc2160, Ecco Data Portal [data set], https://data.nas.nasa.gov/ecco/eccodata/llc_2160/, last access: 1 March 2023b.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L.: Deep Contextualized Word Representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2227–

- 2237, Association for Computational Linguistics, New Orleans, Louisiana, <https://doi.org/10.18653/v1/N18-1202>, 2018.
- Prants, S.: Marine life at Lagrangian fronts, *Prog. Oceanogr.*, 204, 102790, <https://doi.org/10.1016/j.pocean.2022.102790>, 2022.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I.: Improving Language Understanding by Generative Pre-Training, 1–12, 2018.
- Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G.: Daily High-Resolution-Blended Analyses for Sea Surface Temperature, *J. Climate*, 20, 5473–5496, <https://doi.org/10.1175/2007JCLI1824.1>, 2007.
- Seo, H., O'Neill, L. W., Bourassa, M. A., Czaja, A., Drushka, K., Edson, J. B., Fox-Kemper, B., Frenger, I., Gille, S. T., Kirtman, B. P., Minobe, S., Pendergrass, A. G., Renault, L., Roberts, M. J., Schneider, N., Small, R. J., Stoffelen, A., and Wang, Q.: Ocean Mesoscale and Frontal-Scale Ocean–Atmosphere Interactions and Influence on Large-Scale Climate: A Review, *J. Climate*, 36, 1981–2013, <https://doi.org/10.1175/JCLI-D-21-0982.1>, 2023.
- Su, Z., Wang, J., Klein, P., Thompson, A. F., and Menemenlis, D.: Ocean submesoscales as a key component of the global heat budget, *Nat. Commun.*, 9, 775, <https://doi.org/10.1038/s41467-018-02983-w>, 2018.
- Thomas, L. N.: Formation of intrathermocline eddies at ocean fronts by wind-driven destruction of potential vorticity, *Dynam. Atmos. Oceans*, 45, 252–273, 2008.
- Trockman, A. and Kolter, J. Z.: Patches Are All You Need?, arXiv [preprint], <https://doi.org/10.48550/arXiv.2201.09792>, 2022.
- Wang, J., Fu, L.-L., Torres, H. S., Chen, S., Qiu, B., and Menemenlis, D.: On the Spatial Scales to be Resolved by the Surface Water and Ocean Topography Ka-Band Radar Interferometer, *J. Atmos. Ocean. Tech.*, 36, 87–99, 2019.
- Woodson, C. B. and Litvin, S. Y.: Ocean fronts drive marine fishery production and biogeochemical cycling, *P. Natl. Acad. Sci. USA*, 112, 1710–1715, 2015.