https://doi.org/10.5194/egusphere-2025-4847
Preprint. Discussion started: 10 October 2025

Multi-satellite U-Net for high-resolution sea surface temperature reconstruction

Ellin Zhao[1], Edwin Goh[2], Alice Yepremyan[2], Jinbo Wang[3], and Brian Wilson[2]

[1]University of California, Los Angeles
[2]Jet Propulsion Laboratory, California Institute of Technology
[3]Texas A&M University

Correspondence: Ellin Zhao (ellinz@ucla.edu)

Abstract. High-resolution sea surface temperature (SST) products are critical for understanding ocean dynamics at submesoscales (less than tens of kilometers) and their influence on upper ocean physics. While modern infrared (IR) radiometers measure SST at high (∼ 1 km) resolution, they cannot image through clouds, resulting in large gaps in remotely sensed SST. In this study, we address the challenge of reconstructing gap-free high-resolution SST by fusing complementary observations across sensors and time using machine learning (ML). We present the Multi-satellite U-Net for SST Estimation (MUSE), a residual U-Net fuses two days (eight 6-hourly snapshots) of multi-satellite IR and microwave (MW) data into cloud-free SST, and further mosaicked into global SST fields. The MUSE model is trained on 9 months of simulated cloudy SST from the MITgcm LLC4320 1/48° SST product, and evaluated on 2 months of held-out LLC4320 data and out-of-distribution Level 3 satellite data. MUSE outperforms single-time and single-satellite baselines across error, correlation and coherence metrics, achieving a global reconstruction error of 0.035°C on the simulated dataset. On the satellite dataset, MUSE produces results comparable to the state-of-the-art Level 4 MUR 0.01° product. Our results demonstrate the power of ML in synthesizing diverse satellite measurements, each with inherent limitations, into a submesoscale-resolving dataset that enables critical insights into ocean dynamics. Both our data fusion strategy and simulation-to-satellite paradigm can be generalized to other geophysical variables to produce high-resolution, observation-based Earth system fields.

1 Introduction

Sea surface temperature (SST) is an important variable for studying ocean-atmosphere interactions, marine ecosystems, and climatology (O'Carroll et al., 2019). Infrared (IR) radiometers measure SST at high spatial resolutions—as fine as 0.75 km (NOAA/NESDIS/STAR, 2023)— allowing the study of submesoscale processes that impact ocean dynamics (Liu et al., 2017). However, IR radiometers have a major limitation: IR signals cannot penetrate clouds, resulting in gaps in satellite imagery. Since ocean cloud cover can reach up to 68% (Eastman et al., 2011), filling gaps in satellite SST is a critical step in preparing operational high-resolution data products.

Contending with cloud gaps in satellite imagery is a longstanding research problem. Cloud gaps are typically filled by interpolating over temporal data, which leverages the transience of clouds, or multi-satellite data, which leverages the capabilities of different radiometers. Microwave (MW) radiometers are particularly useful because unlike IR devices, they are unaffected by clouds but have coarser resolution (tens of kilometers), capabilities which are the opposite of IR radiometers. Traditional gap-filling methods include kriging (Holdaway, 1996) and Optimal Interpolation (OI) (Reynolds and Smith, 1994; Chin et al., 2017), both of which are statistical methods, and Data INterpolating Empirical Orthogonal Functions (DINEOF) (Alvera-Azcárate et al., 2009), which uses matrix factorization. Although effective and widely used, these methods produce smoothed reconstructions that lack small-scale structures due to a number of factors, including assumptions of linearity and in the case of DINEOF, truncation errors (Barth et al., 2022).

Recent works have used machine learning (ML) to overcome the limitations of traditional methods. Gap-filling of satellite images is fundamentally an image reconstruction task and in computer vision, machine learning models have demonstrated exceptional performance on image reconstruction tasks such as denoising, inpainting, and super-resolution (Archana and Jeevaraj, 2024). Researchers have applied machine learning to fill gaps in SST (Goh et al., 2024; Barth et al., 2022) and other oceanographic variables such as sea surface height (SSH) (Beauchamp et al., 2023; Martin et al., 2023), sea surface salinity (Tian et al., 2022), and Chlorophyll-a (Mehdipour et al., 2025). Machine learning outperforms traditional gap-filling methods because the models can approximate complex nonlinear functions, giving them the capacity to learn complex image priors (Ulyanov et al., 2020), and in the domain of oceanography, complex ocean dynamics.

We provide a review of ML-based methods for SST reconstruction. One of the seminal works in ML-based SST gap-filling is Data INterpolating Convolutional Auto-Encoder (DINCAE) (Barth et al., 2022), which can be interpreted as a generalization of DINEOF. Instead of using matrix factorization as done in DINEOF, DINCAE uses a deep network that can be more comprehensively optimized. The DINCAE model is a U-Net architecture (Ronneberger et al., 2015) that fills gaps using 3 daily snapshots of cloudy IR SST. SST reconstruction has been further improved using vision transformers (ViTs) that enable the integration of mask tokens, making them conducive for gap-filling. One type of ViT is the masked autoencoder (MAE) (He et al., 2022a), which masks random patches of the input to encourage the model to learn descriptive features. The Masked Autoencoder for Sea Surface Temperature Reconstruction under Occlusion (MAESSTRO) model adapts MAE for single-time IR SST gap-filling, and achieves low error for high cloud cover cases (Goh et al., 2024). Zupančič Muc et al. (2025) build on MAESSTRO by using 3 daily IR SST snapshots for their Coarse Reconstruction with ITerative Refinement (CRITER) network that uses a MAE in the coarse reconstruction stage. Other works have foregone IR SST, and instead super resolve cloud-free low-resolution MW data to obtain high-resolution SST. Ducournau and Fablet (2016) use the Super Resolving Convolutional Neural Network (SR-CNN) (Dong et al., 2016) from computer vision to reconstruct high-resolution SST from MW data. Similarly, Fanelli et al. (2024) super-resolve MW SST using a deep network titled dilated Adaptive Deep Residual Network for Super Resolution (dADRSR), which uses channel attention and dilated convolution to further improve the reconstruction quality. Both super-resolution works use single-time SST. A separate group of works have used physics-informed priors to reconstruct SST using ML. Young et al. (2024) reconstruct gap-free IR SST using a temporal-spatial radial basis function neural network (TS-RBFNN) with a Gaussian kernel to emulate the fundamental solution to the heat equation. Despite the integration of physics into the ML model, this method still relies on handcrafted physical priors. Table 1 summarizes the discussed works and provides additional information about the data products used in the studies.

Model	Input SST	Target resolution	Data product(s)	Study area(s)	RMSE (°C)
DINCAE	3 days IR	0.05°	L3 AVHRR	Mediterranean Sea	0.38
CRITER	3 days IR	0.05°, 0.0625°	L3 AVHRR, L3S CNR MED, L3S ODYSSEA	Adriatic Sea, Atlantic Sea, Mediterranean Sea	0.130, 0.391, 0.127
dADRSR	1 day MW	0.01°	L3S CNR MED	Mediterranean Sea	0.31
MAESSTRO	1 day IR	0.02°	LLC4320, LLC2160, L2P VIIRS	Global, Pacific Ocean	0.023, 0.358
MUSE (ours)	2 days (8-time) MW-IR	0.02°	LLC4320, L3S GHRSST NOAA	Global, Mediterranean Sea, Pacific Ocean	**0.035**, **0.15**

Table 1. A summary of machine learning (ML) models for SST reconstruction and their corresponding performance in terms of root mean squared error (RMSE). Our model, Multi-satellite U-net for SST Estimation (MUSE), is distinct for its use of global training and evaluation data, and shows competitive performance on real satellite inputs despite being trained solely on high-resolution simulated SST.

Currently there is a lack of studies that (1) use multi-time *and* multi-satellite SST for gap-filling, and (2) evaluate their models on global SST fields. We fill these research gaps with our Multi-satellite U-net for SST Estimation (MUSE) model that uses multi-time multi-satellite data to reconstruct cloud-free high-resolution (nominally 0.02°) SST. For this study, multi-satellite SST refers to multi-resolution SST captured by IR and MW radiometers. We perform an ablation study on 15 model variants with different inputs (number of time steps, inclusion of MW data) and choose a model that ingests 2 days (eight 6-hourly timesteps) of MW-IR SST as the final configuration for MUSE. Our dataset contains 4 snapshots of SST per day, so we refer to the model using 2 days of MW-IR data as the *8-time MW-IR model*. We show that our MUSE model produces accurate foundation SST, and resolves submesoscale structures. Unlike prior methods, we train and evaluate on *global* SST and conduct a comprehensive evaluation of the model performance in terms of error, correlation and coherence metrics. We also present a preliminary study evaluating our model on real satellite SST that is statistically out-of-distribution (OOD) from the synthetic data used in model training. We compare our satellite SST results to the state-of-the-art Level 4 (L4) MUR 0.01° product (NASA/JPL, 2015) and show that even without model fine-tuning, the MUSE model can overcome the sim-to-real gap (Kadian et al., 2020) and produce reconstructions with RMSEs comparable to prior works. Our approach enables the study of submesoscale processes critical for understanding ocean physics.

2 Datasets

We use a 1/48° global ocean simulation to train our ML model and apply real cloud masks to the simulated SST to emulate realistic satellite data. Our ML model reconstructs SST under the cloud mask, and is trained in a supervised framework using the ground truth (unmasked) SST. SST data near the poles (60-90° S, N) are removed to ignore sea ice and non-uniform gridding effects. In this section, we introduce the data products for model training and evaluation.

2.1 Training dataset

We train our model using the SST fields from a MITgcm global ocean simulation, which is referred to as LLC4320 (Menemenlis et al., 2021). Since its inception in 2016, the simulation has been heavily used for studying submesocale ocean dynamics using machine learning (Su et al., 2018; Goh et al., 2024; Martin et al., 2023). The simulated SST has 1/48° nominal horizontal resolution, and 90 vertical layers on a Lat-Lon-Cap (LLC) grid. Its forcing fields are derived from ERA5 6-hourly winds, air-temperature and pressure, and humidity data. Barotropic tidal forcing was simulated using an equivalent air pressure instead of body force (Arbic et al., 2012). 14-months (September 2011 - November 2012) of hourly snapshots were saved, and sub-sampled to 6-hourly snapshots for this work.

Synthetic gappy IR SST is produced by applying real cloud masks to the ocean simulation output (LLC4320). We derive real cloud masks from the GHRSST NOAA/STAR Level 3 super-collated (L3S) 0.02° dataset (Jonasson et al., 2024), which provides per-pixel quality levels ranging from 1 to 5 where 5 is the best quality, and levels 4-5 denote analysis ready SST. We set pixels with quality level less than 4 as clouds. The daytime and nighttime L3S products are combined to create a dataset with 4 snapshots per day, corresponding to approximately 6-hourly snapshots to match the LLC4320 temporal resolution. We downsample the high-resolution LLC4320 SST by a factor of 12 to simulate MW data at a 0.25° resolution, which corresponds to that of satellite MW SST products. In summary, our training dataset contains cloud-free and cloudy high-resolution (0.02°) IR SST, and cloud-free low-resolution (0.25°) MW SST, all of which are 6-hourly snapshots.

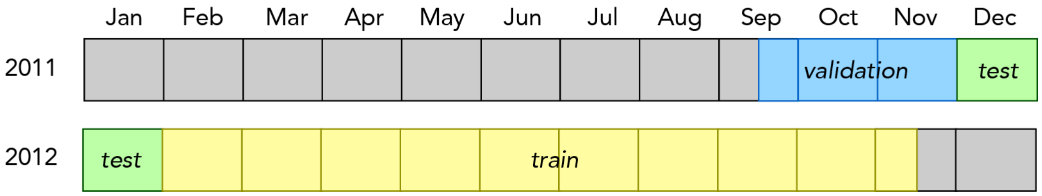

Figure 1. The 14-month LLC4320 simulation (15 September 2011 – 14 November 2012) is split by date into a 9½-month training period (1 February – 14 November 2012), a 2½-month validation period (15 September – 30 November 2011), and a 2-month testing period (1 December 2011 – 31 January 2012). Validation data are used during development to tune hyperparameters and select model checkpoints, whereas the test data—completely held out until final evaluation—provide an unbiased assessment of generalization. The test period deliberately covers the unseen December–January window to assess the model's ability to handle previously unseen seasonal conditions.

2.2 Test dataset

Our MUSE model, trained on simulated cloudy SST, is evaluated on both simulated and real satellite SST datasets to assess its performance and generalization capabilities.

2.2.1 Simulated test data

The primary evaluation is performed on a held-out portion of the simulated LLC4320 dataset. The full 14-month simulation is split by date:

- **Training**: 1 February 2012 - 14 November 2012
- **Validation**: 15 September 2011 - 30 November 2011
- **Testing**: 1 December 2011 - 31 January 2012

This partitioning ensures the test set months are entirely unseen during training and validation, allowing for an unbiased assessment of the model's ability to handle different seasonal conditions. The partitioning of the dataset is illustrated in Figure 1.

2.2.2 Real satellite test data

To evaluate performance in a real-world scenario, we additionally test the model on actual satellite observations from the same testing period.

- **High-resolution infrared (IR) data**: We use the cloudy SST fields from the GHRSST NOAA/STAR L3S 0.02° gridded product (hereafter L3S) as inputs to the model.
- **Low-resolution Microwave (MW) data**: We use the gap-free daily L4 Multi-scale Ultra-high-resolution (MUR) 0.25° gridded SST (NASA/JPL, 2019) as our MW input.

It is important to note that while L3S cloud masks were used to create the *training* data, here we use the actual L3S SST values for *testing*. The strict separation of dates between the training and test sets ensures there is no data leakage.

2.2.3 Preprocessing for sim-to-real transfer

Unlike the clean, simulated data, real satellite SST contains instrument and collation noise, creating an OOD challenge known as the *sim-to-real gap* (Kadian et al., 2020; Hu et al., 2024). To mitigate this, the satellite data undergoes targeted preprocessing:

1. **L4 MUR 0.25° upsampling**: the satellite MW data is bi-linearly upsampled across time to match the 6-hourly temporal resolution of the L3S data.
2. **L3S denoising:** Extreme outliers (outside 4 standard deviations of the mean) are removed from the L3S data. A 2×2 pixel median filter is then applied to reduce measurement noise, which coarsens the effective resolution from 0.02° to 0.04°.
3. **Bias correction:** An additive bias between the upsampled MUR data and the L3S data, caused by the temporal interpolation, is calculated and removed from the MW data.

These preprocessing steps proved sufficient for this initial investigation of the sim-to-real gap.

https://doi.org/10.5194/egusphere-2025-4847
Preprint. Discussion started: 10 October 2025

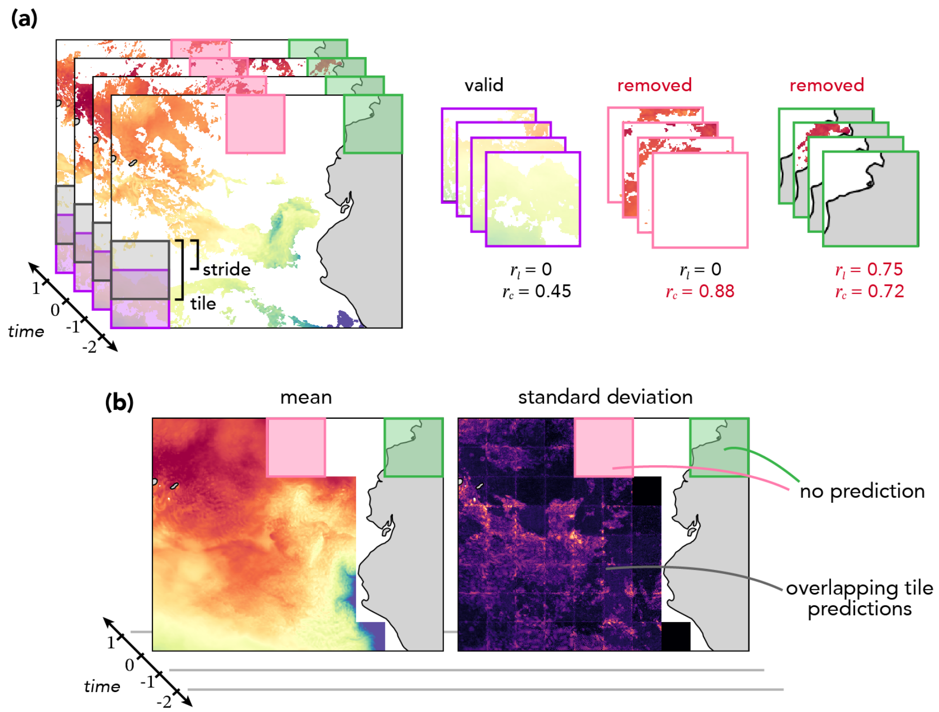

Figure 2. Cloudy SST is divided into tiles that are input to our machine learning model, and the reconstructed tiles are combined to create gap-free global SST. In this example, the tile size is $4 \times 128 \times 128$ and the tile stride is $1 \times 64 \times 64$. **(a)** A 4-time input is divided into tiles, and tiles with too much land or cloud are excluded from reconstruction. **(b)** The model reconstructs the middle time ($t = 0$) of the input, and spatially overlapping tiles are aggregated by calculating the mean and standard deviation. White regions indicate areas where no model prediction is made (e.g., too cloudy or too much land).

3 MUSE methodology

3.1 Machine learning model inputs

The global SST fields are divided into smaller tiles before being passed to the model. This windowing process is parametrized by the tile size and the stride between tiles, as shown in Figure 2(a). We choose the spatial tile size to be 128×128 pixels, and the temporal size is one of [1, 4, 8, 12]. The ground truth SST ($\mathbf{T}$), cloudy infrared SST ($\mathbf{T}_\mathrm{c}$), and cloud-free microwave SST ($\mathbf{T}_\mu$) are all tiled in the same manner.

For a given tile, the domain is segmented into binary masks representing land ($\mathbf{M}_\mathrm{l}$), visible ocean ($\mathbf{M}_\mathrm{vis}$) and cloudy ocean ($\mathbf{M}_\mathrm{c}$) areas. Based on these masks, the cloud ratio, r_c, and land ratio, r_l, are defined as:

$$r_c = \frac{||\mathbf{M}_\mathrm{c}||_0}{||\mathbf{M}_\mathrm{vis} + \mathbf{M}_\mathrm{c}||_0}, \tag{1}$$

$$r_l = \frac{||\mathbf{M}_\mathrm{l}||_0}{||\mathbf{M}_\mathrm{vis} + \mathbf{M}_\mathrm{c} + \mathbf{M}_\mathrm{l}||_0}, \tag{2}$$

Variable	Description
$\mathbf{T}$	Ground truth high-resolution SST
$\hat{\mathbf{T}}$	Estimated cloud-free high-resolution SST
$\mathbf{T}_\mathrm{c}$	Cloudy high-resolution (IR) SST
$\mathbf{T}_\mu$	Cloud-free low-resolution (microwave) SST
$\lvert\nabla_{xy}\mathbf{T}\rvert$	Magnitude of spatial SST gradient
$\lvert\nabla_t\mathbf{T}\rvert$	Magnitude of temporal SST gradient
$\mathbf{M}_\mathrm{c}$	Cloud mask
$\mathbf{M}_\mathrm{l}$	Land mask
$\mathbf{M}_\mathrm{vis}$	Visible/ clear mask

Table 2. Summary of notation used for the SST and cloud mask variables.

where $||\cdot||_0$ is the ℓ_0-pseudo norm (the count of non-zero elements). Tiles with excessive land or cloud cover ($r_c > 0.75$, $r_l > 0.1$) are removed from the dataset, and no prediction is made for these tiles. Examples of valid and removed tiles are shown in Figure 2(a). Our best model uses 2 days of SST as input, resulting in a tile size of $8 \times 128 \times 128$, and the stride is chosen to be $4 \times 128 \times 128$. The raw training dataset contains 1.8 million tiles with 624,569 usable tiles after selection based on r_c and r_l. The dataset tiles are normalized by the training SST mean (21.59°C) and standard deviation (2.94°C).

After normalization, the cloud and land regions of each tile are initialized. For a tile with cloudy IR SST, $\mathbf{T}_\mathrm{c}$, and land, cloud and visible masks, $\mathbf{M}_\mathrm{l}$, $\mathbf{M}_\mathrm{c}$, $\mathbf{M}_\mathrm{vis}$, the model input is:

$$-2.5 \cdot \mathbf{M}_\mathrm{l} + \mathbf{C} \odot \mathbf{M}_\mathrm{c} + \mathbf{T}_\mathrm{c} \odot \mathbf{M}_\mathrm{vis}, \tag{3}$$

where $\odot$ is the Hadamard product, and $\mathbf{C}$ is the pixel-wise cloud fill value. The land initialization value, -2.5, is constant, and is chosen to be different than valid SST. Since the SST is normalized to be zero-mean and unit-standard deviation, the value -2.5 is at the tail end of the training SST data, and is therefore distinct from most valid SST. We test two choices of the cloud fill value, $\mathbf{C}$. We use the mean of the visible SST, $\bar{T}_\mathrm{vis}$, as a per-tile constant fill value:

$$\bar{T}_\mathrm{vis} = \frac{1}{||\mathbf{M}_\mathrm{vis}||_0} \sum \mathbf{T}_\mathrm{c} \odot \mathbf{M}_\mathrm{vis}, \tag{4}$$

where the summation is over all values. When available, we use the microwave SST ($\mathbf{T}_\mu$) as the cloud fill value. In summary, the cloud fill values depend on the satellite data used:

$$\mathbf{C} = \begin{cases} \bar{T}_\mathrm{vis}, & \text{single-satellite} \\ \mathbf{T}_\mu, & \text{multi-satellite.} \end{cases} \tag{5}$$

Based on the dataset tiling parameters (stride, size), the reconstructed tiles may overlap. In those cases, we aggregate the predictions by calculating the mean and standard deviation, which act as our prediction and uncertainty estimate. An example of overlapping tile predictions is shown in Figure 2(b).

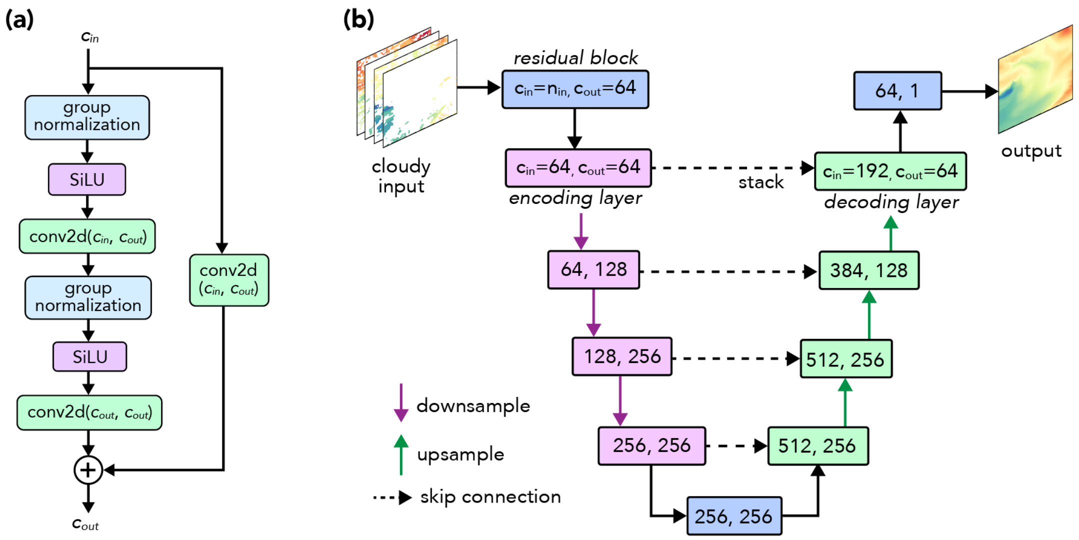

Figure 3. The U-Net reconstruction model architecture. **(a)** The residual block takes in a tensor with c_{in} input channels and outputs a tensor with c_{out} output channels. The $\texttt{conv2d}(c_{\text{in}}, c_{\text{out}})$ operator has c_{in} input channels and c_{out} output channels, and performs a 2D spatial convolution with a 3×3 kernel. **(b)** The U-Net model takes in n_{in} timesteps of cloudy SST and outputs a cloud-free estimate of the middle timestep. The architecture contains residual blocks (blue), encoding layers (purple) and decoding layers (green). The numbers for each layer indicate the input and output channel dimensions.

3.2 SST reconstruction neural network

Our MUSE model is based on a modern U-Net architecture (Gupta and Brandstetter, 2022), which is a convolutional autoencoder model that has been extensively used for image reconstruction tasks across many fields including remote sensing (Ibtehaz and Rahman, 2020; He et al., 2022b; Hou et al., 2021).

The MUSE model is built from a series of *residual blocks*, and the residual block outputs the sum of two paths. The first path performs the following sequence of operations twice: group normalization (with a group size of 32), sigmoid linear unit (SiLU) activation, and 2D convolution (with kernel size of 3×3). The second path is a residual connection that prevents vanishing gradients during optimization (He et al., 2016). Figure 3(a) summarizes the operations in a residual block. The layers in the U-Net model consist of residual blocks. Each encoding and decoding layer contains 2 residual blocks with input channel size, c_{in}, and output channel size, c_{out}. The encoding layers are followed by $2\times$ downsampling using a learned convolution, and analogously, the decoding layers are followed by $2\times$ upsampling using a learned transposed convolution. The overall U-Net structure is shown in Figure 3(b), and our final U-Net model, which uses an 8-time MW-IR input, has 23.6 million parameters.

The model takes in n timesteps of cloudy SST, and reconstructs the middle timestep. It is trained to minimize the difference between the model estimate ($\hat{\mathbf{T}}$) and the ground truth gap-free SST ($\mathbf{T}$). The difference is evaluated using the mean squared error (MSE) between $\hat{\mathbf{T}}$ and $\mathbf{T}$. MSE-trained models are known to produce smoothed results because outliers are amplified through the squaring operation (Zhao et al., 2017), and we mitigate this by adding a loss term using the spatial gradient magnitude to penalize inaccurate small-scale reconstructions. The spatial gradient magnitude of an input $\mathbf{T}$ is:

$$|\nabla_{xy}\mathbf{T}| = \sqrt{\nabla_x \mathbf{T}^2 + \nabla_y \mathbf{T}^2}. \tag{6}$$

The full training loss is an affine combination of the MSE and spatial gradient MSE, and γ_1 and γ_2 are weights for the two loss components:

$$L(\mathbf{T},\hat{\mathbf{T}}) = \gamma_1 \mathrm{MSE}(\mathbf{T},\hat{\mathbf{T}}) + \gamma_2 \mathrm{MSE}(|\nabla_{xy}\mathbf{T}|,|\nabla_{xy}\hat{\mathbf{T}}|). \tag{7}$$

We set $\gamma_1 = 1$ and $\gamma_2 = 5$ so that the SST and SST gradient losses have similar magnitudes during training. SST input tiles are grouped into mini-batches of size 72 for training.

The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The learning rate follows a one-cycle scheduler (Smith and Topin, 2019) and increases from $1\mathrm{e}^{-5}$ to the maximum value of $5\mathrm{e}^{-4}$ over the first 18% of the total 200 training epochs. Our models were trained using 4 NVIDIA A100 GPUs, and took 36 hours to converge.

3.3 Evaluating model performance

We comprehensively evaluate our reconstructions using multiple metrics including the root mean squared errors (RMSEs) of the SST itself ($\mathbf{T}$), the spatial gradient magnitude ($|\nabla_{xy}\mathbf{T}|$), and the temporal gradient magnitude ($|\nabla_t\mathbf{T}|$) of the SST. We also report the Pearson correlation coefficient, which is calculated over smaller regional domains and then averaged globally. This ensures that the correlation is not dominated by the large range of global SST values. We report the correlation coefficients for SST and the small-scale SST, which are denoted r and r_{ss} respectively. Small-scale SST is derived by subtracting the low-resolution MW data from the SST, resulting in a high-pass filtered SST anomaly.

We use coherence to measure model performance over a range of spatial scales. The coherence between the ground truth, $\mathbf{T}$, and reconstruction, $\hat{\mathbf{T}}$, is defined as:

$$c(k) = \frac{|\mathrm{CSD}_{\hat{\mathbf{T}},\mathbf{T}}(k)|^2}{\mathrm{PSD}_{\hat{\mathbf{T}}}(k)\mathrm{PSD}_{\mathbf{T}}(k)}, \tag{8}$$

where $\mathrm{CSD}_{\hat{\mathbf{T}},\mathbf{T}}$ is the cross spectral density between $\hat{\mathbf{T}}$ and $\mathbf{T}$, and $\mathrm{PSD}_{\hat{\mathbf{T}}}$ and $\mathrm{PSD}_{\mathbf{T}}$ are the power spectral densities of $\hat{\mathbf{T}}$ and $\mathbf{T}$ respectively. Coherence ranges from 0 to 1, and higher values are better. To compare our reconstructions to the cloudy input, we define the temporally-averaged input cloud ratio:

$$\frac{1}{n}\sum_{i=-\frac{n}{2}}^{\frac{n}{2}-1} \mathbf{M}_{\mathrm{c}}(x,y,t=i), \tag{9}$$

where n is the number of timesteps in the input, and $t = 0$ is the target timestep that is reconstructed by our MUSE model. The input cloud ratio shows how often each pixel is visible over the temporal domain of the input.

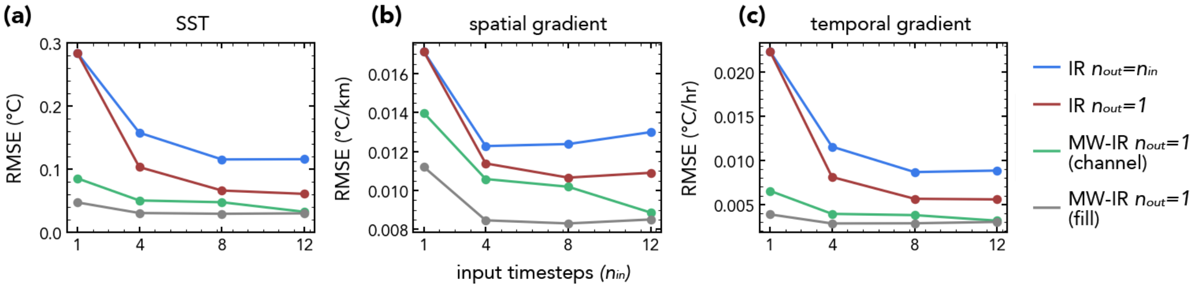

Figure 4. Results from an ablation study on the LLC4320 validation set, evaluating how performance is affected by the number of input timesteps (n_{in}), output timesteps (n_{out}), and the use of single-satellite (IR) versus multi-satellite (MW-IR) data. Models were assessed on their global RMSE for **(a)** SST, **(b)** spatial gradient, and **(c)** temporal gradient. The best-performing model, which achieves low errors across all three metrics, uses an 8-timestep MW-IR input to reconstruct a single frame (n_{in}=8, n_{out}=1; gray). Note that for n_{in}=1, the n_{out}=n_{in}case (blue) is equivalent to the n_{out}=1 case (red).

4 Results and discussion

The MUSE model accurately reconstructs SST under clouds as measured by error, correlation and coherence metrics. We demonstrate MUSE's ability to reconstruct large-scale and submesoscale SST through a series of regional reconstructions. We summarize the LLC4320 reconstruction performance through global metrics, and present an initial study on the reconstruction of OOD satellite SST.

4.1 Ablation study on LLC4320 validation set

Across 15 different model configurations, our best performing model is the one that uses an 8-time MW-IR input. In this section, we present an ablation study of these 15 models, which vary in the choice of model input (n_{in}) and output timesteps (n_{out}), and the incorporation of multi-satellite input data. The global LLC4320 validation dataset is used for this ablation study. The performance of these 15 models in terms of (a) SST, (b) spatial gradient, and (c) temporal gradient RMSE is shown in Figure 4.

First, we consider the effect of n_{out}. For varying n_{in}, we compare single-satellite (IR) models where $n_{\text{out}} = n_{\text{in}}$ (blue line) and $n_{\text{out}} = 1$ (red). $n_{\text{out}} = 1$ models reconstruct only the middle timestep of the input. The middle timestep is maximally correlated to all input timesteps, so reconstructing the middle timestep is easier than reconstructing all input timesteps. As such, for all values of n_{in}, the $n_{\text{out}} = 1$ models achieve a lower RMSE. All subsequent models in this section use $n_{\text{out}} = 1$.

Next, we investigate the impact of incorporating low-resolution MW data by comparing the single-satellite (IR) model against two multi-satellite (MW-IR) configurations. We test two methods for incorporating the MW data:

1. **MW Channel**: IR and MW data are stacked as separate input channels, and cloud gaps in IR data are filled using mean visible temperature, $\bar{T}_{\text{vis}}$, defined in Equation (4).

2. **MW Fill**: MW data is used to fill cloud gaps in the IR input, as described in Equation (5).

The inclusion of MW data significantly reduces reconstruction RMSE compared to the IR-only, single-satellite model, as shown by the red, green, and gray lines in Figure 4. The results also show that the MW fill method outperforms the MW channel method across all metrics. While both multi-satellite inputs provide the same core information, the MW fill method is more compact. This streamlined input representation likely makes it easier for the model to learn dependencies between the large-scale MW context and the fine-scale IR features, leading to better performance.

Finally, we consider the effect of n_{in}. We vary n_{in} from 1 to 12, which corresponds to single-time to 3 days. The RMSE of the SST, spatial gradient and temporal gradient decreases as n_{in} increases, but the performance plateaus at $n_{\text{in}} = 8$. This is true for both the single- and multi-satellite models. For the MW-IR fill model (gray line), the RMSE begins to rise when n_{in} is increased from 8 to 12, and this effect is most prominent for the spatial gradient RMSE. Models using long ($n_{\text{in}} \geq 12$) input time series appear prone to overfitting, especially for small-spatial scale features. Based on these results, our best model uses $n_{\text{out}} = 1$, $n_{\text{in}} = 8$, and incorporates MW data using the MW fill method. In the following sections, we will show results from both our best model (8-time MW-IR) and its single-satellite counterpart (8-time IR) to demonstrate the specific cases where MW data improves reconstruction.

4.2 Regional LLC4320 analysis: Indian Ocean

4.2.1 Multi-time model input leverages cloud transience

The multi-satellite model outperforms the single-satellite model in reconstructing a highly cloudy region of the Indian Ocean (9.35-14.35° S, 107.25-112.25° E). The target reconstruction is 3 December 2011 at 12:00:00 UTC. The 8-time model input is centered around the target time ($t = 0$), and includes SST from 2 December 12:00:00 UTC to 4 December 18:00:00 UTC, and this 8-time MW-IR input is shown in Figure 5(a). The target time is highly cloudy ($r_c = 0.83$), which makes a single-time reconstruction impossible. The multi-time input has a lower cloud ratio ($r_c = 0.52$), but improves the spatial coverage of visible SST by providing additional information over time as shown in Figure 5(b).

4.2.2 Multi-satellite model improves eddy reconstruction

We compare our reconstructions to the ground truth cloud-free SST at $t = 0$. Overall, both the IR and MW-IR models perform well, achieving low RMSEs of 0.104°C and 0.039°C respectively. However, the MW-IR model performs better in certain regions. This is evident in Figure 5(c), which shows a comparison of the IR and MW-IR reconstructions in the top row and the reconstruction errors (target - reconstruction) in the bottom row. The IR reconstruction has local regions of temperature bias where the SST is consistently hotter or colder than the ground truth. Temperature bias appears in regions that remain continuously cloudy over both space and time (i.e., white/lighter regions in the bottom half of Figure 5(b)), leaving few reference points for model interpolation. This is evident in the north-western corner of the domain where the target SST contains a cold core eddy. Over the eddy, the cloud ratio is high ($r_c = 0.78$) causing the IR model to overestimate the temperature by 0.5°C. Since the eddy is approximately 50 km wide, its structure is visible in the MW input, and as a result, the MW-IR

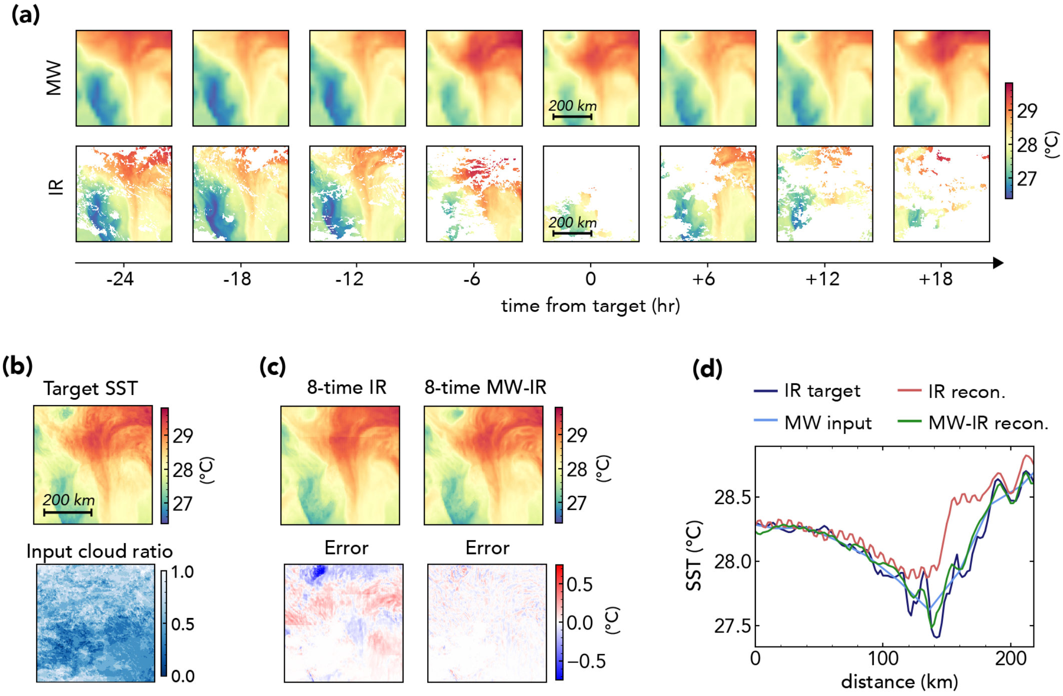

Figure 5. Our 8-time IR and MW-IR models produce accurate reconstructions of a highly cloudy region (500×500 km) region of the Indian Ocean (9.35-14.35° S, 107.25-112.25° E). **(a)** The model input is 8 timesteps of MW and IR SST, centered around the target reconstruction time, $t = 0$. **(b)** Ground truth SST for $t = 0$ (top), and the the input cloud ratio over time (bottom). **(c)** The 8-time IR and MW-IR reconstructions (top), and the reconstruction errors, $\mathbf{T} - \hat{\mathbf{T}}$ (bottom). **(d)** Cross section of a cold core eddy from the target SST (black), MW input (blue), IR reconstruction (red) and the MW-IR reconstruction (green).

reconstruction is highly accurate. In general the MW-IR reconstruction exhibits no temperature bias because the MW input always provides large-scale SST structures.

250 The longitudinal cross-sections of the target (dark blue line), MW input (light blue), IR reconstruction (red), and MW-IR reconstruction (green) across the eddy are shown in Figure 5(d). This plot further verifies that the MW-IR model retains the large-scale MW input structure, and further recovers small-scale details synthesized from the IR data. The cross-section also shows a slight oscillating artifact in the IR reconstruction, which can be easily removed with Gaussian blurring. Although there are limitations to the single-satellite model, it still produces a realistic result due to the use of 8-time input data, and may be 255 useful when MW data is unavailable.

4.3 Regional LLC4320 analysis: Pacific Ocean

In this section, we show the temporal and spatial reconstruction quality of our MUSE model on a region of the Pacific Ocean near California. The study area, which captures part of the California current, is delimited by 31-37° N, 123-129° W.

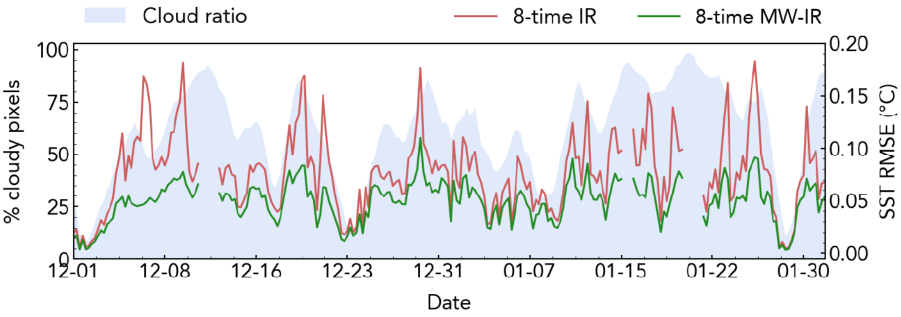

Figure 6. The low temporal reconstruction error for our Pacific Ocean study area indicates that our models can generalize to months that are unseen during training and that the incorporation of MW information consistently outperforms an IR-only scenario. On the left y-axis, we show the cloud ratio (shaded blue area), and on the right y-axis, we show the RMSE for the 8-time IR (red) and MW-IR (green) models. Gaps in the RMSE correspond to days that are too cloudy, and no model prediction is made.

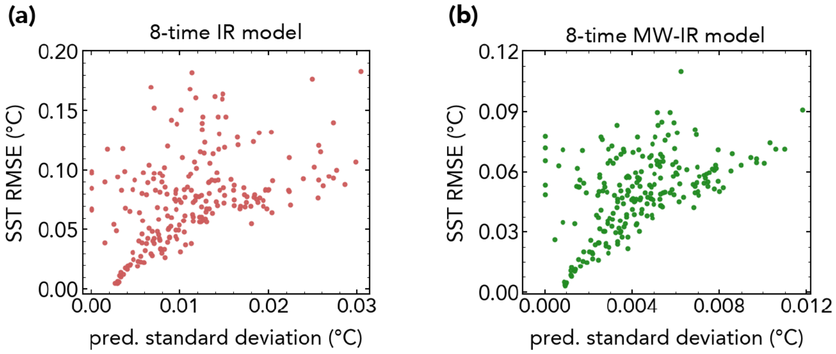

Figure 7. Model uncertainty can be measured using the standard deviation of predictions generated from the overlapping portions of each tile. This is useful for assessing reconstruction quality when no ground truth SST is available. The prediction standard deviation is positively correlated with RMSE for both the **(a)** 8-time IR and **(b)** 8-time MW-IR models.

4.3.1 Model generalization across time

The MUSE model is accurate in predicting temporal SST values, and exhibits no bias in reconstructing unseen months. For our study area, we reconstruct the SST from 1 December 2011 to 30 January 2012, and calculate the regional RMSE for each timestep. Figure 6 shows the RMSE time series for the IR model (red line) and the MW-IR model (green line), as well as the cloud ratio time series (shaded blue area). When the cloud ratio is too high, no model prediction is made and no RMSE is reported (i.e. 12 December). The months reconstructed in this example do not appear in the training data. Our results show no temporal bias, indicating that our model can generalize to unseen seasonal traits. While the RMSE does not exhibit a temporal trend, it does follow the cloud ratio. Since the cloud ratio is independent of the reconstruction, it can be used to quantify the reconstruction quality and uncertainty when a ground truth reference is unavailable.

https://doi.org/10.5194/egusphere-2025-4847
Preprint. Discussion started: 10 October 2025

4.3.2 Quantifying model uncertainty

Our reconstruction uncertainty can be quantified using values derived from the model input and output. In the previous section, Figure 6 showed a correlation between RMSE and cloud ratio. Here, we quantify the relationship between RMSE and cloud ratio using the Spearman correlation coefficient (ρ), and introduce another variable for measuring model uncertainty. There is high correlation between cloud ratio and RMSE: for the IR model, $\rho = 0.703$, and for the MW-IR model, $\rho = 0.745$. Reported values are statistically significant ($p \approx 0$).

There is also positive correlation between the RMSE and prediction standard deviation. The standard deviation is calculated over the ensembled tile predictions, which is discussed in Section 2. We plot the prediction standard deviation against the RMSE for the (a) IR and (b) MW-IR models in Figure 7. In the plots, each point represents a single timestep in our test set date range. For both models, the standard deviation and RMSE are positively correlated according to ρ. The IR model has a correlation of $\rho = 0.437$, and the MW-IR model has a correlation of $\rho = 0.453$. The positive correlation indicates that the model prediction has more uncertainty when the prediction is inaccurate, which is the expected behavior. A combination of the cloud ratio and prediction standard deviation can be used to estimate model uncertainty since both variables are positively correlated with error, and do not require a ground truth reference.

4.3.3 Multi-satellite model improves small-scale reconstruction

Our 8-time MW-IR model accurately reconstructs small-spatial scale SST, as demonstrated by the spatial gradient error and the spectral coherence values. We focus on a single timestep of the test set on 9 January 2012 00:00:00 UTC. Figure 8(a) shows the cloudy SST, cloud-free target, and the reconstruction from our MW-IR model. Figure 8(b) shows the spatial gradient, $|\nabla_{xy}\mathbf{T}|$, of these SST fields to demonstrate the model's ability to reconstruct submesoscale structures. For brevity, we do not show the IR model reconstruction. The SST at the target time is moderately cloudy ($r_c = 0.43$), and the cloudy areas in the target are visible at other times as shown by the input cloud ratio in Figure 8(c). The SST and SST gradient reconstructions are highly accurate, and there is no discernible difference even in a closeup of a 200×200 km region of the gradient field in Figure 8(b). To elucidate the errors in our reconstruction, we show the gradient error field (reconstruction - ground truth) in Figure 8(d). The errors are small compared to the gradient magnitude, and the areas with poor gradient reconstruction correspond to highly cloudy regions in the input. Over this study area, the SST RMSE is 0.037°C, and the gradient SST RMSE is 0.012 °C km^{-1}, indicating that our reconstruction captures large- and small-scale features under regions of contiguous clouds.

To assess model performance over a range of spatial scales, we show the spectral coherence, $c(k)$, between the ground truth and reconstruction. Two sub-regions—boxes i and ii are shown in Figure 9(a)—that have high cloud cover are chosen for analysis. Box i has a cloud ratio of 0.63, and box ii has a cloud ratio of 0.96. For both sub-regions, we show the spectral coherence averaged over the latitudinal and longitudinal directions in Figure 9(b). The coherence of the IR and MW-IR reconstructions are shown in red and green, respectively. For reference, we indicate the resolution of the MW input data in gray. For both boxes, the coherence is high for large-scale SST (small k), and around the MW resolution, the coherence begins to drop. For smaller

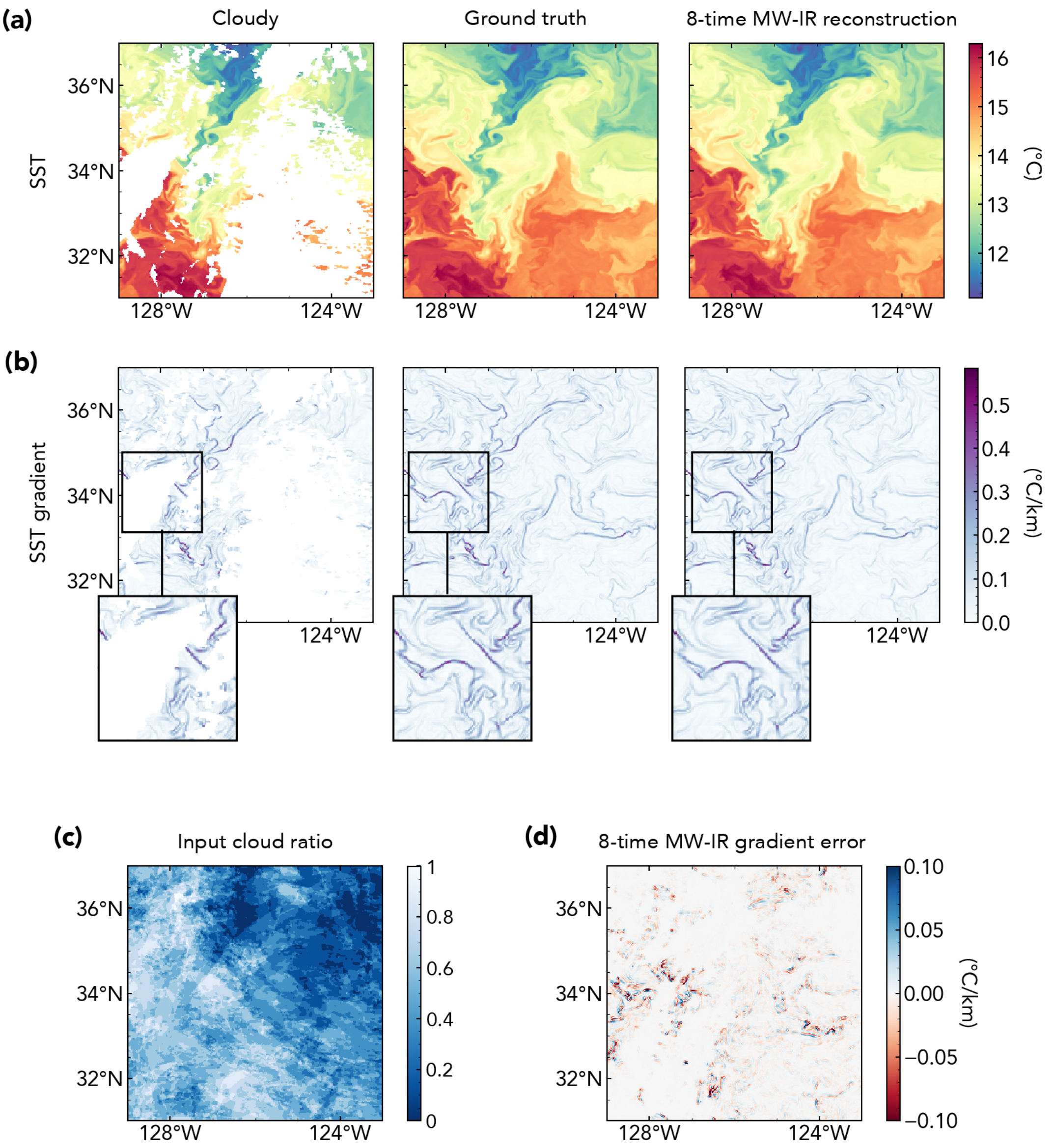

Figure 8. SST and spatial gradient reconstruction for the Pacific Ocean study area on 9 January 2012. **(a)** Cloudy SST, cloud-free SST and our 8-time MW-IR reconstruction. **(b)** Cloudy gradient SST, cloud-free gradient SST and our 8-time MW-IR gradient reconstruction. **(c)** Cloudiness over the 8-time input. **(d)** Gradient reconstruction error.

300 scales (large k), the MW-IR coherence is broadly greater than the IR coherence. This demonstrates that the MW-IR model has a super-resolving ability because it improves reconstruction at wavelengths smaller than the input (i.e. sub-MW wavelengths).

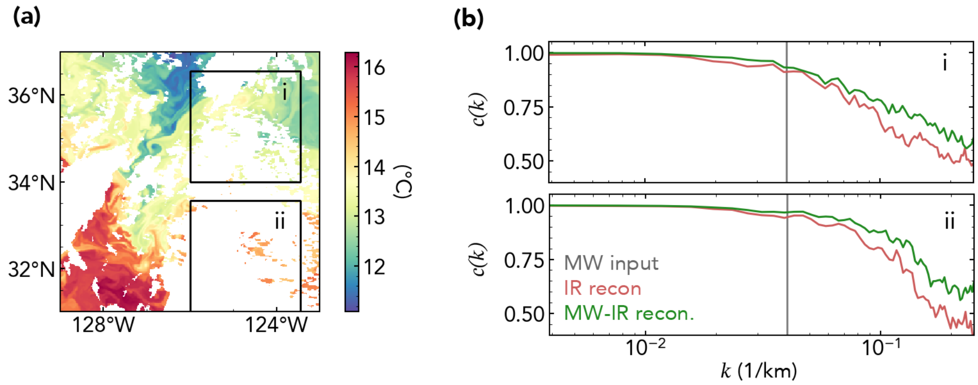

Figure 9. The spectral coherence of the 8-time IR and MW-IR model reconstructions for the Pacific Ocean study area on 9 January 2012 indicates that our MW-IR model outperforms the IR model in small-scale reconstruction. **(a)** We show the cloudy SST for the study area, and two sub-regions (boxes i, ii) that are chosen for the coherence analysis. **(b)** Spectral coherence for box i (top) and box ii (bottom). The MW data resolution ($k = 0.04$ km^{-1}) is shown in gray, and the IR and MW-IR reconstruction coherence are shown in red and green respectively.

Table 3. The global performance of our 8-time IR and MW-IR models averaged over the LLC4320 test set. The RMSE of the SST ($\mathbf{T}$) spatial gradient, ($|\nabla_{xy}\mathbf{T}|$) and temporal gradient, ($|\nabla_t\mathbf{T}|$) are reported. Pearson correlation coefficient at all spatial scales (r) and at small-scales (sub-25 km, r_{ss}) are reported. For RMSE, lower values (↓) are better, and for correlation, higher values (↑) are better.

	RMSE (↓)			Correlation (↑)	
Model	$\mathbf{T}$ (°C)	$\lvert\nabla_{xy}\mathbf{T}\rvert$ (°C km^{-1})	$\lvert\nabla_t\mathbf{T}\rvert$ (°C h^{-1})	All scales (r)	Small-scale (r_{ss})
8-time IR	0.080	0.012	0.006	0.986	0.855
8-time MW-IR	0.035	0.010	0.003	0.996	0.950

4.4 Global LLC4320 SST reconstruction

We complete our discussion of the LLC4320 test set by reporting the global error and correlation of our model reconstructions. We report the RMSE of the SST ($\mathbf{T}$), spatial gradient magnitude ($|\nabla_{xy}\mathbf{T}|$), and temporal gradient magnitude ($|\nabla_t\mathbf{T}|$) for the 8-time IR and MW-IR models. These values are listed in Table 3. The MW-IR model achieves a remarkably low SST RMSE of 0.035°C, and across the three metrics, the MW-IR model achieves a lower error than the IR model. The smallest RMSE reduction from the IR to MW-IR model occurs for the spatial gradient, $|\nabla_{xy}\mathbf{T}|$. The MW-IR model reduces the $|\nabla_{xy}\mathbf{T}|$ RMSE by 0.002 °C km^{-1}, which is equivalent to a 20% reduction of the IR RMSE. The tempered reduction of the gradient RMSE is expected because the MW data provides only large-scale (> 25 km) information, but it is notable that the use of MW data improves the small-scale reconstruction at all.

In addition to RMSE, we report the correlation between $\mathbf{T}$ and $\hat{\mathbf{T}}$ for all spatial scales and small (< 25 km) spatial scales in Table 3. For the correlation at all spatial scales (r), both models achieve nearly perfect correlation (IR model $r = 0.986$, MW-IR

https://doi.org/10.5194/egusphere-2025-4847
Preprint. Discussion started: 10 October 2025

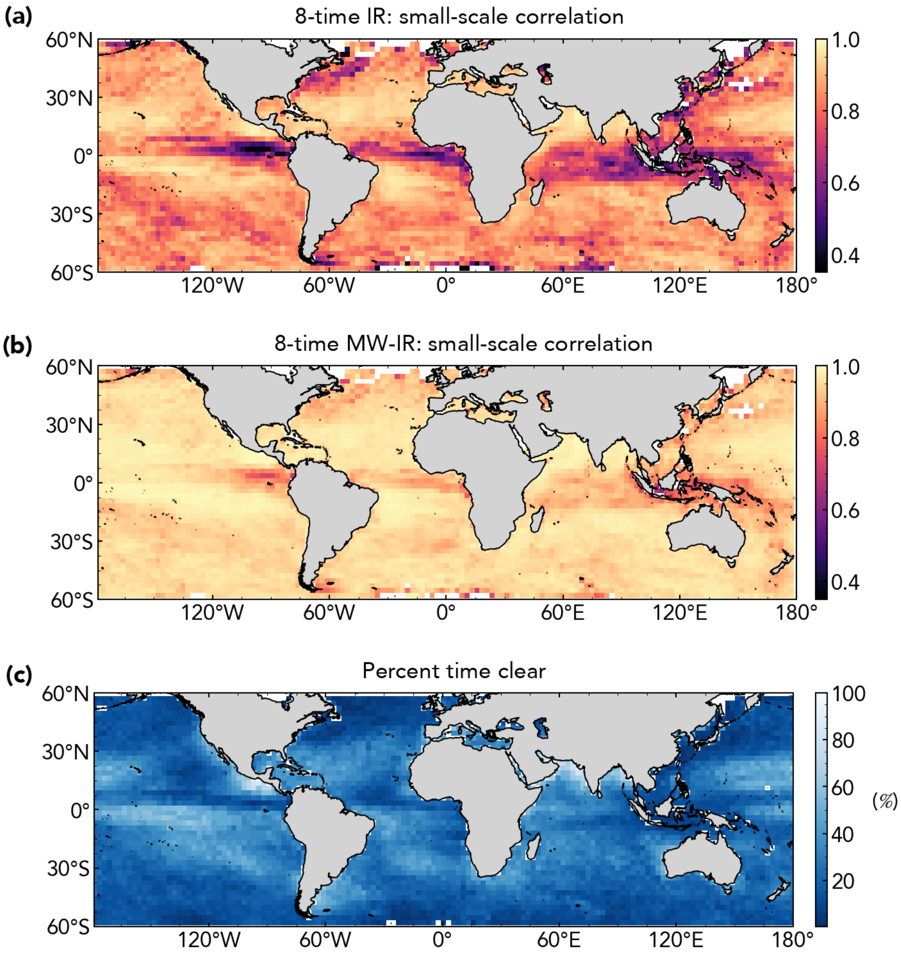

Figure 10. Global small-scale correlation maps for our **(a)** 8-time IR and **(b)** MW-IR models. SST at high latitudes are excluded from our dataset. The global cloud ratio across the test set dates is shown in **(c)**.

model $r = 0.996$). The small-scale correlation provides a more interesting comparison. For the small-scale correlation (r_{ss}), the MW-IR model has a high correlation of 0.950, while the IR model's correlation drops to 0.855. The correlation is still high 315 for both models, but the MW-IR model clearly performs better than the IR model for small-scale SST. In our regional examples, we showed how the MW-IR model improved small-scale reconstruction, and the global RMSE and correlation values affirm that this result holds in general.

To verify that there is no regional bias in model performance, we show the global small-scale correlation maps, averaged over time, for the IR and MW-IR models in Figure 10(a) and (b) respectively. Areas with no valid data—near the poles where 320 sea ice is present—are indicated with white pixels. We additionally show the percent of time that a pixel is clear in Figure 10(c). Note that this is the opposite of the cloud ratio. The global maps show that the spatial correlation pattern follows that of the cloudiness, and there is no other regional bias in correlation (i.e. elevated error near coastlines).

Table 4. The performance of our 8-time MW-IR model on 2 regions of the L3S SST. The RMSE of the SST ($\mathbf{T}$) spatial gradient, ($|\nabla_{xy}\mathbf{T}|$) and temporal gradient, ($|\nabla_t\mathbf{T}|$) are reported. Pearson correlation coefficient at all spatial scales (r) and at small-scales (sub-25 km, r_{ss}) are reported. For RMSE, lower values (↓) are better, and for correlation, higher values (↑) are better.

	RMSE (↓)			Correlation (↑)	
Region	$\mathbf{T}$ (°C)	$\lvert\nabla_{xy}\mathbf{T}\rvert$ (°C km^{-1})	$\lvert\nabla_t\mathbf{T}\rvert$ (°C h^{-1})	All scales (r)	Small-scale (r_{ss})
Pacific Ocean	0.178	0.038	0.017	0.954	0.688
Mediterranean Sea	0.133	0.027	0.015	0.953	0.787
Mean	0.156	0.033	0.016	0.934	0.738

4.5 Preliminary study on L3S SST reconstruction

We perform a preliminary analysis of the MUSE model on real satellite data, and show that despite the sim-to-real gap, our results are still competitive with prior works. We evaluate our best performing model, the 8-time MW-IR model, on two regions—the Northern Pacific Ocean (30-38° N, 120-130° W), and the Mediterranean Sea (36.5-44.5° N, 2-12° E). Our model uses cloudy L3S 0.02° SST and cloud-free L4 MUR 0.25° SST as model inputs instead of simulated SST. The full dataset details are in Section 2.

We quantitatively evaluate our model by masking visible regions of the L3S SST, providing us with ground truth cloud-free SST. For each timestep, the expanded cloud mask is the union of the original and vertically flipped mask. We report the L3S dataset RMSE and correlation values for the Pacific Ocean and Mediterranean Sea in Table 4.

4.5.1 Measuring the sim-to-real gap

We compare the averaged regional L3S results in Table 4 to the global LLC4320 results in Table 3 to measure the sim-to-real gap. This is a rough comparison because the LLC4320 results are global. Furthermore, roughly 12% more pixels are masked out in the real data to obtain reconstructions over visible pixels that can be used to calculate RMSE, which makes the reconstruction problem more challenging. As expected, the performance for the L3S dataset is worse because it is out-of-distribution (OOD). The LLC4320 RMSE is 0.035°C, while the L3S error is 0.155°C. For the SST, spatial gradient, and temporal gradient RMSEs, the LLC4320 error is $4.4\times$, $3.2\times$, and $5.3\times$ smaller than the L3S error. For the small-scale correlation, the LLC4320 value is $1.3\times$ larger than the L3S value. The difference between the correlation at all scales is negligible because the values are very high (> 0.95). In Section 2, we discussed how the L3S SST is OOD from the LLC4320 SST due to noise. Despite the existence of a sim-to-real gap, the results on L3S SST are promising because we only use preprocessing to mitigate the OOD effect. Our model, which is trained on unprecedented volumes of high-resolution, global *simulated* SST can serve as a good function for finetuning on *real* data where ground truth validation samples are sparse.

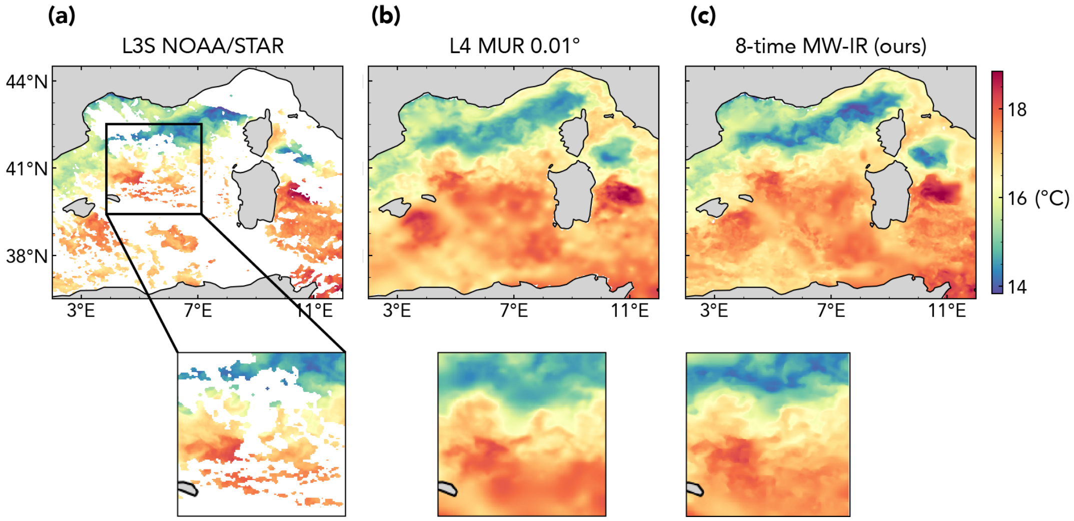

Figure 11. Reconstruction of Mediterranean Sea (Sea of Sardinia) SST. **(a)** Cloudy L3S SST, **(b)** cloud-free L4 MUR 0.01° SST, and **(c)** our ML-based cloud-free reconstruction. The inset shows region where our reconstruction is more detailed than the L4 MUR SST.

4.5.2 Comparison to prior works

345 We compare our L3S results to prior works, and show that our 8-time MW-IR model is competitive with the works listed in Table 1. This is a rough comparison because the prior works use different data products, and have varying target resolutions (0.01-0.0625°). The DINCAE (Barth et al., 2022), dADRSR (Fanelli et al., 2024), and CRITER (Zupančič Muc et al., 2025) models evaluate regions of the Mediterranean Sea, and achieve RMSEs of 0.38°C, 0.31°C, and 0.13°C respectively. Despite being trained only on simulation data, our 8-time MW-IR model has an RMSE of 0.13°C for the Mediterranean Sea, which is the same as CRITER. Our work focuses on optimizing the input data to the model, and out of all the discussed works, our inputs are the most comprehensive (8-time multi-satellite). Prior works focus on adapting new model architectures for SST reconstruction, while our architecture is a relatively simple U-Net. By combining the strengths of our work and prior works, a multi-time multi-satellite ML model can achieve unprecedented accuracy in reconstructing SST.

4.5.3 Comparison to L4 MUR 0.01° product

We compare our reconstructions to the L4 MUR 0.01° SST product (NASA/JPL, 2015). For brevity, we refer to the L4 MUR 0.01° SST as L4 SST. In these examples, we do not expand the cloud masks as we did in the previous sections. We visually compare our reconstruction to the L4 SST because even in the non-cloudy regions, the L4 SST does not match the L3S SST. This is visible in Figure 11, where we show the cloudy L3S, L4, and our reconstructed SST over the Mediterranean Sea. At 42.5° N, 7° E, the L4 SST is hotter than the L3S SST, which is contrary to our work that considers L3S SST to be the ground truth. In the cloudy regions, the L4 SST is blurred, while our reconstruction is not, and this is especially evident in the inset shown in Figure 11.

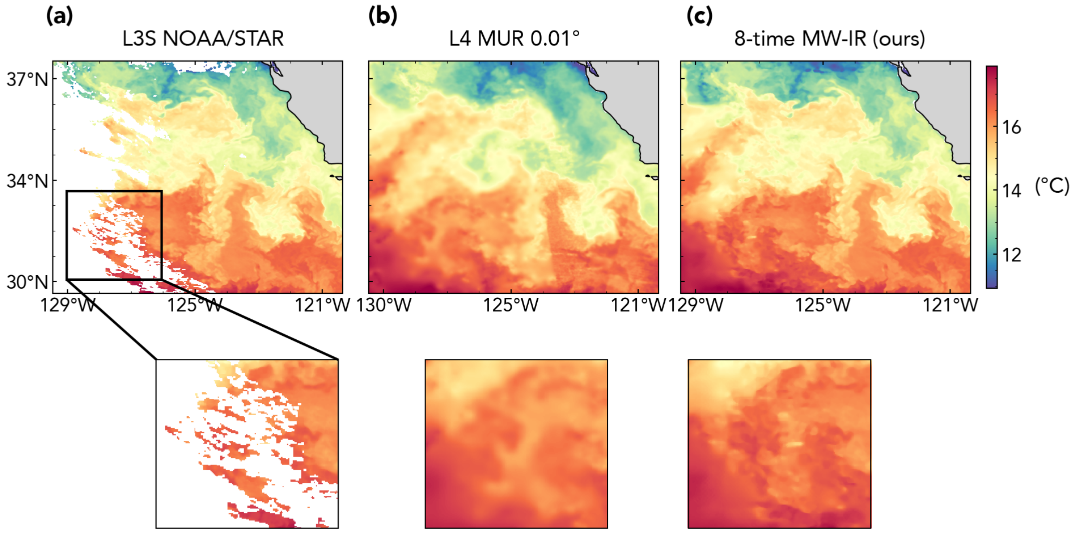

Figure 12. Reconstruction of Pacific Ocean SST. **(a)** Cloudy L3S SST, **(b)** cloud-free L4 MUR 0.01° SST, and **(c)** our ML-based cloud-free reconstruction. The inset shows region where our reconstruction is more detailed than the L4 MUR SST.

For the Pacific Ocean region shown in Figure 12, our gap-free SST appears more detailed than the L4 SST. In this example, the clouds only affect the western region of the study area, but the L4 SST and our reconstruction still differ in the visible region around 31° N, 122° W. The inset shows that the L4 SST and our reconstruction agree in the large-scale SST, but our model reconstructs finer details. In the middle of the inset, a small region of SST does not appear realistic and may be from noise at other input timesteps. Since we lack ground truth SST, our comparisons to the L4 MUR 0.01° SST are strictly qualitative, and only assess the existence and realism of the small-scale features. Section 4.5.1 provides quantitative measures of our reconstruction quality. Our initial study on L3S data is promising because minimal steps are added to mitigate the sim-to-real gap, but our results are still comparable to both prior works and L4 MUR 0.01° SST.

5 Conclusions

High-resolution SST is important for the study of submesoscale ocean dynamics. While IR radiometers can measure at fine resolutions, IR data contains cloud gaps, which severely limits its use. Traditional methods of gap-filling reduce the effective resolution of the data, obscuring small-scale SST dynamics. In this work, we present a ML-based method for reconstructing high-resolution gap-free SST from multi-time multi-satellite data. Our MUSE model performs well in both large- and small-scale reconstructions, which we verify using RMSE, correlation and spectral coherence values. Our model is trained on a dataset of global SST, providing the model with a diverse set of ocean dynamics. This is possible because we use simulated SST, allowing us to have ground truth cloud-free SST over the entire ocean. Despite training solely on simulated SST, MUSE also performs well in reconstructing real satellite data.

We evaluate our LLC4320-trained model on L3S SST, and use data preprocessing to mitigate the sim-to-real gap. Future work can expand this study, and fine-tune the model weights on a smaller dataset of L3S SST pairs. The model architecture can also be improved by incorporating channel attention: this may be particularly impactful because our model uses a long (8-timestep) temporal input. In our work, we assume that the microwave data is cloud-free. For this reason, we use a L4 MW SST product that is already gap-filled for our satellite data evaluation. In future works, we can loosen our assumption and allow MW data to contain gaps, so that we can use a L3 MW product, such as the L3U Advanced Scanning Microwave Radiometer (AMSR) data, to avoid incurring errors from gap-filled inputs. If MW SST is assumed to have gaps, there may be some regions with no visible SST information. To combat this, geographical information (latitude, longitude) can be included in the model input to ensure that every pixel has some information.

Multi-variable data can be used to further bolster the model input. High-resolution (~ 1 km) sea surface height (SSH) captured by the Surface Water and Ocean Topography (SWOT) mission can be used together with SST to produce dynamically consistent high-resolution data products for monitoring and studying submesoscale upper ocean processes and their influence. This study has demonstrated the ability of machine learning models to synthesize multi-time multi-satellite SST into accurate and detailed cloud-free SST. Our work shows that global gap-free high-resolution SST, which is integral for understanding the Earth's climate, is achievable using ML in the near future.

Code and data availability. All used datasets are publicly available on the JPL Physical Oceanography Distributed Active Archive Center (PO.DAAC) website (NASA/JPL, 2015, 2019; Jonasson et al., 2024). We provide full instructions for data preparation, model training and evaluation in our GitHub repository (Zhao, 2025a). In addition to the code, samples of downloaded and processed data, along with model weights, are publicly available on the Harvard Dataverse (Zhao, 2025b).

Author contributions. EG prepared the Level 3 data for model training and analysis. EG and AY developed the infrastructure for distributed model training, and implemented the Residual U-Net architecture. EZ, EG and AY designed the model experiments and determined model configurations for the ablation study. EZ generated tables and plots for the final model, and drafted the manuscript. JW conceptualized the problem, and advised on the interpretation of model performance. 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 authors declare that they have no conflict of interest.

Acknowledgements. A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM001 8D0004) and was funded by the Advanced Information Systems Technology (AIST) program. The High Performance Computing resources used in this investigation were provided by funding from the JPL Information and Technology Solutions Directorate. E. Zhao is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2034835. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. J. Wang is funded by NASA grant 80NSSC25K7643. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper. The authors thank Ed Armstrong for helpful discussions about data products.

References

Alvera-Azcárate, A., Barth, A., Sirjacobs, D., and Beckers, J.-M.: Enhancing temporal correlations in EOF expansions for the reconstruction of missing data using DINEOF, Ocean Science, 5, 475–485, https://doi.org/10.5194/os-5-475-2009, 2009.

Arbic, B., Richman, J., Shriver, J., Timko, P., Metzger, J., and Wallcraft, A.: Global Modeling of Internal Tides Within an Eddying Ocean General Circulation Model, Oceanography, 25, 20–29, https://doi.org/10.5670/oceanog.2012.38, 2012.

Archana, R. and Jeevaraj, P. S. E.: Deep learning models for digital image processing: a review, Artificial Intelligence Review, 57, https://doi.org/10.1007/s10462-023-10631-z, 2024.

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, Geoscientific Model Development, 15, 2183–2196, https://doi.org/10.5194/gmd-15-2183-2022, 2022.

Beauchamp, M., Febvre, Q., Georgenthum, H., and Fablet, R.: 4DVarNet-SSH: end-to-end learning of variational interpolation schemes for nadir and wide-swath satellite altimetry, Geoscientific Model Development, 16, 2119–2147, https://doi.org/10.5194/gmd-16-2119-2023, 2023.

Chin, T. M., Vazquez-Cuervo, J., and Armstrong, E. M.: A multi-scale high-resolution analysis of global sea surface temperature, Remote Sensing of Environment, 200, 154–169, https://doi.org/10.1016/j.rse.2017.07.029, 2017.

Dong, C., Loy, C. C., He, K., and Tang, X.: Image Super-Resolution Using Deep Convolutional Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 295–307, https://doi.org/10.1109/tpami.2015.2439281, 2016.

Ducournau, A. and Fablet, R.: Deep learning for ocean remote sensing: an application of convolutional neural networks for super-resolution on satellite-derived SST data, in: 2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS), p. 1–6, IEEE, https://doi.org/10.1109/prrs.2016.7867019, 2016.

Eastman, R., Warren, S. G., and Hahn, C. J.: Variations in Cloud Cover and Cloud Types over the Ocean from Surface Observations, 1954–2008, Journal of Climate, 24, 5914–5934, https://doi.org/10.1175/2011jcli3972.1, 2011.

Fanelli, C., Ciani, D., Pisano, A., and Buongiorno Nardelli, B.: Deep learning for the super resolution of Mediterranean sea surface temperature fields, Ocean Science, 20, 1035–1050, https://doi.org/10.5194/os-20-1035-2024, 2024.

Goh, E., Yepremyan, A., Wang, J., and Wilson, B.: MAESSTRO: Masked Autoencoders for Sea Surface Temperature Reconstruction under Occlusion, Ocean Science, 20, 1309–1323, https://doi.org/10.5194/os-20-1309-2024, 2024.

Gupta, J. K. and Brandstetter, J.: Towards Multi-spatiotemporal-scale Generalized PDE Modeling, https://arxiv.org/abs/2209.15616, 2022.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., and Girshick, R.: Masked Autoencoders Are Scalable Vision Learners, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, https://doi.org/10.1109/cvpr52688.2022.01553, 2022a.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., and Xue, Y.: Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–15, https://doi.org/10.1109/tgrs.2022.3144165, 2022b.

Holdaway, M.: Spatial modeling and interpolation of monthly temperature using kriging, Climate Research, 6, 215–225, https://doi.org/10.3354/cr006215, 1996.

Hou, Y., Liu, Z., Zhang, T., and Li, Y.: C-UNet: Complement UNet for Remote Sensing Road Extraction, Sensors, 21, 2153, https://doi.org/10.3390/s21062153, 2021.

https://doi.org/10.5194/egusphere-2025-4847
Preprint. Discussion started: 10 October 2025

Hu, X., Li, S., Huang, T., Tang, B., Huai, R., and Chen, L.: How Simulation Helps Autonomous Driving: A Survey of Sim2real, Digital Twins, and Parallel Intelligence, IEEE Transactions on Intelligent Vehicles, 9, 593–612, https://doi.org/10.1109/tiv.2023.3312777, 2024.

Ibtehaz, N. and Rahman, M. S.: MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Networks, 121, 74–87, https://doi.org/10.1016/j.neunet.2019.08.025, 2020.

Jonasson, O., Gladkova, I., Ignatov, A., and Kihai, Y.: GHRSST NOAA/STAR ACSPO v2.81 0.02 degree L3S Daily Dataset from LEO Satellites (GDS version 2), https://doi.org/10.25921/D8EF-8597, 2024.

Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., and Batra, D.: Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance?, IEEE Robotics and Automation Letters, 5, 6670–6677, https://doi.org/10.1109/lra.2020.3013848, 2020.

Liu, J., Emery, W., Wu, X., Li, M., Li, C., and Zhang, L.: Computing Coastal Ocean Surface Currents from MODIS and VIIRS Satellite Imagery, Remote Sensing, 9, 1083, https://doi.org/10.3390/rs9101083, 2017.

Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, in: International Conference on Learning Representations, https://openreview.net/forum?id=Bkg6RiCqY7, 2019.

Martin, S. A., Manucharyan, G. E., and Klein, P.: Synthesizing Sea Surface Temperature and Satellite Altimetry Observations Using Deep Learning Improves the Accuracy and Resolution of Gridded Sea Surface Height Anomalies, Journal of Advances in Modeling Earth Systems, 15, https://doi.org/10.1029/2022ms003589, 2023.

Mehdipour, E., Xi, H., Barth, A., Alvera-Azcárate, A., Wilhelm, A., and Bracher, A.: Assessment of gap-filling techniques applied to satellite phytoplankton composition products for the Atlantic Ocean, https://doi.org/10.5194/egusphere-2025-112, 2025.

Menemenlis, D., Hill, C., Henze, C. E., Wang, J., and Fenty, I.: Southern Ocean Pre-SWOT Level-4 Hourly MITgcm LLC4320 Native Grid 2km Oceanographic Dataset Version 1.0, https://doi.org/10.5067/PRESW-ASJ10, 2021.

NASA/JPL: GHRSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (v4.1), https://doi.org/10.5067/GHGMR-4FJ04, 2015.

NASA/JPL: GHRSST Level 4 MUR 0.25deg Global Foundation Sea Surface Temperature Analysis (v4.2), https://doi.org/10.5067/GHM25-4FJ42, 2019.

NOAA/NESDIS/STAR: GHRSST Level 2P NOAA ACSPO SST v2.80 from VIIRS on NOAA-21 Satellite, https://doi.org/10.5067/GHN21-2P280, 2023.

O'Carroll, A. G., Armstrong, E. M., Beggs, H. M., Bouali, M., Casey, K. S., Corlett, G. K., Dash, P., Donlon, C. J., Gentemann, C. L., Høyer, J. L., Ignatov, A., Kabobah, K., Kachi, M., Kurihara, Y., Karagali, I., Maturi, E., Merchant, C. J., Marullo, S., Minnett, P. J., Pennybacker, M., Ramakrishnan, B., Ramsankaran, R., Santoleri, R., Sunder, S., Saux Picart, S., Vázquez-Cuervo, J., and Wimmer, W.: Observational Needs of Sea Surface Temperature, Frontiers in Marine Science, 6, https://doi.org/10.3389/fmars.2019.00420, 2019.

Reynolds, R. W. and Smith, T. M.: Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation, Journal of Climate, 7, 929–948, https://doi.org/10.1175/1520-0442(1994)007<0929:igssta>2.0.co;2, 1994.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, p. 234–241, Springer International Publishing, ISBN 9783319245744, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Smith, L. N. and Topin, N.: Super-convergence: very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, edited by Pham, T., p. 36, SPIE, https://doi.org/10.1117/12.2520589, 2019.

Su, Z., Wang, J., Klein, P., Thompson, A., and Menemenlis, D.: Ocean submesoscales as a key component of the global heat budget, Nature Communications, https://doi.org/10.1038/s41467-018-02983-w, 2018.

Tian, T., Cheng, L., Wang, G., Abraham, J., Wei, W., Ren, S., Zhu, J., Song, J., and Leng, H.: Reconstructing ocean subsurface salinity at high resolution using a machine learning approach, Earth System Science Data, 14, 5037–5060, https://doi.org/10.5194/essd-14-5037-2022, 2022.

Ulyanov, D., Vedaldi, A., and Lempitsky, V.: Deep Image Prior, International Journal of Computer Vision, 128, 1867–1888, https://doi.org/10.1007/s11263-020-01303-4, 2020.

Young, C.-C., Cheng, Y.-C., Lee, M.-A., and Wu, J.-H.: Accurate reconstruction of satellite-derived SST under cloud and cloud-free areas using a physically-informed machine learning approach, Remote Sensing of Environment, 313, 114 339, https://doi.org/10.1016/j.rse.2024.114339, 2024.

Zhao, E. Q. L.: ellinzhao/muse: Code release for Zenodo, https://doi.org/10.5281/ZENODO.16056414, 2025a.

Zhao, E. Q. L.: Data for the paper "Multi-time multi-satellite reconstruction of gap-free high-resolution sea surface temperature", https://doi.org/10.7910/DVN/7QFSOA, 2025b.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J.: Loss Functions for Image Restoration With Neural Networks, IEEE Transactions on Computational Imaging, 3, 47–57, https://doi.org/10.1109/tci.2016.2644865, 2017.

Zupančič Muc, M., Zavrtanik, V., Barth, A., Alvera-Azcarate, A., Ličer, M., and Kristan, M.: CRITER 1.0: A coarse reconstruction with iterative refinement network for sparse spatio-temporal satellite data, https://doi.org/10.5194/gmd-2024-208, https://gmd.copernicus.org/preprints/gmd-2024-208/, 2025.