DIRECT 1.0: A diffusion-based generative model for dense sea surface temperature reconstructions from sparse satellite observations

Grega Rovšček[1], Matjaž Ličer[2,3,5], Alexander Barth[4], and Matej Kristan[1]

[1]Faculty of Computer and Information Science, Visual Cognitive Systems Lab, University of Ljubljana, Ljubljana, Slovenia
[2]Slovenian Environment Agency, Office for Meteorology, Hydrology and Oceanography, Ljubljana, Slovenia
[3]National Institute of Biology, Marine Biology Station, Piran, Slovenia
[4]Department of Astrophysics, Geophysics and Oceanography, Geohydrodynamics and Environment Research, University of Liège, Liège, Belgium
[5]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

Correspondence: Grega Rovšček (grega.rovscek@fri.uni-lj.si)

Abstract. Satellite sea surface temperature (SST) observations are frequently obscured by cloud cover, creating large gaps that must be reconstructed for many oceanographic and climate applications. Because multiple high-resolution SST fields may be consistent with the same sparse observations, this reconstruction problem is inherently ambiguous. Nevertheless, most existing approaches remain deterministic, producing a single estimate that is often overly smooth, may contain unrealistic artifacts, and provides limited or unreliable uncertainty estimates. To address these limitations, we introduce DIRECT, a conditional generative framework for dense SST reconstruction that models the full distribution of plausible solutions rather than a single deterministic estimate. DIRECT is based on a rectified flow-matching formulation, conditioned on temporal context and day-of-year seasonality, and trained on sparse Level-3 observations. It presents an observation-guided rectification that anchors the generative trajectory to measured pixels at every integration step. By sampling multiple reconstructions, DIRECT produces an ensemble of physically plausible SST fields, enabling both an accurate mean reconstruction and spatially resolved uncertainty estimates. The latter is adjusted with a simple post-hoc variance term to avoid under-dispersed uncertainty estimates. Experiments on three Level-3 SST datasets (Mediterranean, Adriatic, and Atlantic) show that DIRECT sets a new state-of-the-art, reducing Root Mean Square Error by 6–14 % compared with the strongest published method, while better preserving mesoscale structure. Further analysis of spatial scale correlations indicates that DIRECT maintains physically consistent textures even when reconstructing large, completely unobserved regions. Performance improvements remain robust across a wide range of cloud-coverage conditions, enabling reliable SST reconstruction from sparse satellite observations.

1 Introduction

Sea surface temperature (SST) is a key variable at the ocean–atmosphere interface, regulating exchanges of heat, moisture, and momentum and influencing atmospheric convection, storm development, and large-scale climate variability such as ENSO and monsoon systems (Rahman and Rahaman, 2024; Li et al., 2024; Garcia-Soto et al., 2021; Ricchi et al., 2023). SST is therefore recognized as an essential climate variable (Bojinski et al., 2014) and is widely used in climate monitoring, numerical weather prediction, ocean reanalysis, and marine ecosystem studies (Senatore et al., 2020; Donlon et al., 2007).

Global SST observations are primarily obtained from satellite-based infrared (IR) and microwave radiometers on low-Earth-orbit and geostationary platforms (O'Carroll et al., 2019). IR instruments provide fine spatial detail, but their usable coverage is strongly limited by clouds. At a given time, cloud systems can hide large coherent areas of the ocean surface, resulting in structured gaps that persist across space and time. These gaps present a challenge for downstream applications that require spatially complete SST fields (climatological datasets, surface boundary conditions for atmospheric models, etc.).

Early SST reconstruction methods based on statistical interpolation and low-rank decompositions (e.g., Optimal Interpolation, EOFs (Taburet et al., 2019; Alvera-Azcárate et al., 2005)) recover large-scale variability but are limited in representing fine-scale and nonlinear ocean dynamics, due to linearity and stationarity assumptions. Recent learning-based methods have substantially improved reconstruction quality by exploiting spatial and temporal correlations in the data. Convolutional autoencoders such as DINCAE/DINCAE2 (Barth et al., 2020, 2022) provide uncertainty surrogates but often produce overly smooth fields and underestimate mesoscale variability. Transformer-based architectures, including MAESSTRO (Goh et al., 2024) and CRITER (Zupančič Muc et al., 2025), improve the modeling of long-range dependencies but remain fundamentally deterministic, producing a single best-guess reconstruction that struggles to capture ambiguity under heavy cloud cover, complicates uncertainty estimation, and may introduce unrealistic structured artifacts (Figure 1, bottom-right).

In machine learning, particularly in computer vision, diffusion-style generative models have achieved state-of-the-art performance in image restoration by learning full conditional distributions rather than point estimates (Croitoru et al., 2023; Li et al., 2023; Xing et al., 2024). Pixel re-injection as in RePaint (Lugmayr et al., 2022) and subsequent refinements (Liu et al., 2024) has been successfully exploited to enforce observation consistency during sampling for inpainting. However, these methods typically operate on single images and do not address spatiotemporally structured gaps, physical masks (e.g., land), or calibrated uncertainty in geophysical fields. A growing body of work applies diffusion-style generative models to Earth-science reconstruction tasks. Barth et al. (Barth et al., 2024) demonstrate diffusion-based gap-filling for satellite ocean color (chlorophyll-a), producing ensembles and evaluating uncertainty reliability. For SST specifically, CARE-SST (Choo et al., 2025) applies a denoising diffusion probabilistic model (DDPM (Ho et al., 2020)) style approach with historical context to reconstruct cloud-contaminated SST. Related efforts also extend diffusion-based reconstruction to broader ocean-temperature settings using observation-guided sampling and simulation pretraining (Song et al., 2025). In atmospheric data assimilation, physics-guided diffusion frameworks incorporate physical regularization to improve coherence under sparse observations (Wang et al., 2025). Two concurrent studies extend this direction to global, climate-scale reconstruction: Qian et al. (2026) pretrain a spatiotemporal diffusion prior on climate-model simulations and reanalysis to reconstruct monthly historical temperature and precipitation fields from sparse instrumental records, while Li et al. (2026) perform diffusion posterior sampling to fuse coarse-resolution microwave retrievals with sparse in situ measurements into global ensemble SST estimates. While these works establish promise of generative modeling for geosciences, a crucial difference to the work presented in this paper is that they operate on a single shot or coarse temporal context (Choo et al., 2025; Qian et al., 2026), rely on guidance schemes without explicit temporal

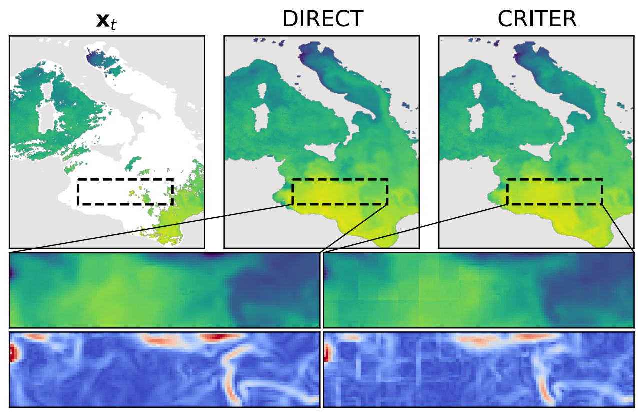

Figure 1. DIRECT reconstructs spatially coherent temperature fields from sparse observation inputs and preserves smooth mesoscale structure in reconstructed regions. DIRECT further addresses the problem of physically inconsistent reconstructions – notice that the blocky high-frequency gradients emerging in CRITER ((Zupančič Muc et al., 2025), bottom right), which are absent in DIRECT reconstructions (bottom left).

encoding (Song et al., 2025; Li et al., 2026), require gap-free training data (Qian et al., 2026; Choo et al., 2025) and do not target calibrated per-pixel uncertainty for dense spatiotemporal reconstruction (Wang et al., 2025).

In this study we propose DIRECT, a diffusion-inspired generative model for dense SST reconstruction from partially observed sparse satellite measurements. DIRECT formulates SST gap-filling as a *conditional flow-matching* task (Lipman et al., 2023), where a deterministic probability flow transforms noise realizations into physically consistent SST fields conditioned on sparse measurements, temporal context, and seasonal information. The model integrates (i) a backbone equipped with SST-specific global conditioning, (ii) a spatio-temporal context representation that preserves the temporal origin of auxiliary information, and (iii) an observation-guided rectification mechanism that maintains consistency with measured pixels throughout the reconstruction process. By sampling multiple reconstructions, DIRECT produces both an accurate mean estimate and spatially resolved uncertainty, which is further automatically calibrated to avoid ensemble collapse. The main contributions are:

- We formulate dense SST reconstruction as a *conditional flow-matching* problem, integrating spatiotemporal context, seasonal conditioning, and hard observation consistency within a unified end-to-end generative model.
- We develop a *temporal offset encoding* strategy for auxiliary context frames.
- We design a lightweight *post-hoc uncertainty calibration* procedure that compensates for ensemble under-dispersion and improves the reliability of per-pixel uncertainty estimates without altering the training objective.

A preliminary version of this work was presented in (Rovšček et al., 2026), which the present work extends in several ways. The reconstruction method has been revised following extended evaluation under operationally realistic conditions, resulting in a simpler and more robust formulation. The experimental evaluation is broadened to include an additional state-of-the-art baseline (CARE-SST; (Choo et al., 2025)), a stratified performance analysis across low, moderate, and high cloud-coverage regimes, and a probabilistic assessment of the reconstructed ensemble via the Continuous Ranked Probability Score. Additionally, a series of new ablation studies examines the individual contributions of seasonal context, mask-aware conditioning, temporal encoding design, and future context availability for near-real-time operational applications.

2 DIRECT architecture and implementation

We begin by introducing key notations and definitions. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_T\} = \{\mathbf{x}_t\}_{t=1}^T$ denote a sequence of SST measurements $\mathbf{x}_t \in \mathbb{R}^{W\times H}$, and $\mathbf{M} = \{\mathbf{m}_t\}_{t=1}^T$ a sequence of corresponding binary masks $\mathbf{m}_t \in \{0,1\}^{W\times H}$, with ones indicating valid observations, and let $\mathbf{m}_l$ be a constant binary mask, with zeros indicating land. Our task is to recover a sequence of dense (gap-free) reconstructed fields $\{\boldsymbol{\mu}_t\}_{t=1}^T$ along with per-pixel uncertainty estimates $\Sigma = \{\boldsymbol{\sigma}_t\}_{t=1}^T$. Here, W denotes the width and H the height, in pixels, of the SST frame.

Reconstructing SST fields is fundamentally a spatio-temporal inference task because ocean surface temperatures evolve continuously under physical constraints. Consequently, measurements acquired close in time to day t are typically informative for recovering missing values, whereas observations from more distant times tend to contribute less. Previous work (Barth et al., 2022; Zupančič Muc et al., 2025) showed that most relevant temporal information is contained within a three-day window centered at the target frame. Following these findings, we use observations from days $t-1$, t, and $t+1$ when reconstructing missing regions at time t.

Because adjacent context frames may themselves be partially observed, we densify the temporal context by opportunistically replacing missing pixels with observations from nearby days. For a pixel that is missing at time $t-1$, we use the closest available observation from $t-2$ or $t-3$. Analogously, missing pixels at $t+1$ are filled from $t+2$ or $t+3$. The resulting filled context frames are denoted as $\mathbf{x}'_{t-1}$ and $\mathbf{x}'_{t+1}$. Note that this search depth $\Delta_{\text{filled}} = 3$ is distinct from the three-day context window, and extends the total temporal footprint of the input to days $t-3$ through $t+3$. A value drawn from an earlier day substitutes for the state at $t-1$ and stays informative only over the timescale on which SST remains temporally autocorrelated, so the same short temporal decorrelation that motivates a narrow context window also bounds the depth over which substitution is meaningful. To retain the temporal origin of each filled value, we associate every context frame with a one-hot encoded mask. For the past context, $\mathbf{M}_{t-1} \in \{0,1\}^{3\times W\times H}$ encodes whether a pixel originates from day $t-1$, $t-2$, or $t-3$. For the future context, $\mathbf{M}_{t+1} \in \{0,1\}^{3\times W\times H}$ similarly encodes offsets $t+1$, $t+2$, or $t+3$. Since each pixel is filled from the temporally nearest valid frame, these channels are mutually exclusive, while a null vector $[0,0,0]$ denotes a pixel that remains invalid (land or persistent cloud cover) after the search. The central frame is represented by a two-channel mask $\mathbf{M}_t \in \{0,1\}^{2\times W\times H}$, which distinguishes observed from missing (or land) pixels at time t. We define the temporal context fields as $\mathbf{C}_{1t} = [\mathbf{x}'_{t-1}, \mathbf{x}_t, \mathbf{x}'_{t+1}]$ and the corresponding masks as $\mathbf{C}_{2t} = [\mathbf{M}_{t-1}, \mathbf{M}_t, \mathbf{M}_{t+1}]$, whose construction for a general search depth is summarized in Algorithm 1. Together, these form the complete spatio-temporal context $\mathbf{C}_t = (\mathbf{C}_{1t}, \mathbf{C}_{2t})$. Following good practices of prior

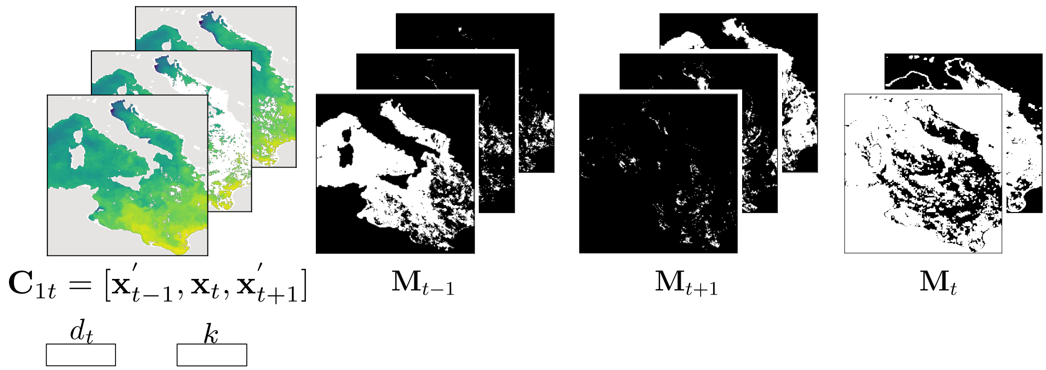

Figure 2. Context information fields provided to DIRECT at input. DIRECT uses a three-day SST window $\mathbf{C}_{1t} = [\mathbf{x}'_{t-1}, \mathbf{x}_t, \mathbf{x}'_{t+1}]$, with corresponding one-hot encoded masks $\mathbf{M}_{t-1}$ and $\mathbf{M}_{t+1}$ indicating the temporal origin of filled pixels, and $\mathbf{M}_t$ marking observed versus missing values in the central frame. Additional conditioning is provided by the day-of-year (d_t) and the flow timestep k.

work (Barth et al., 2022; Zupančič Muc et al., 2025), we additionally include the day-of-year (DoY) d_t as seasonal context information. Figure 2 shows an example of context information.

Algorithm 1 Temporal context construction for target day t. The accumulator $\mathbf{f}$ marks pixels that have already been assigned a temporal origin, and $\mathbf{s}_j$ marks pixels filled from offset j. Because $\mathbf{s}_j$ is nonzero only where $\mathbf{f} = 0$, the origin channels are mutually exclusive by construction.

Input: SST sequence $\mathbf{X}$, mask sequence $\mathbf{M}$, target day t, search depth Δ_{filled}

Output: context fields $\mathbf{C}_{1t}$, temporal-origin masks $\mathbf{C}_{2t}$

Past context:
$\mathbf{x}'_{t-1} \leftarrow \mathbf{0}, \quad \mathbf{f} \leftarrow \mathbf{0}$
for $j = 1$ **to** Δ_{filled} **do**
  $\mathbf{s}_j \leftarrow \mathbf{m}_{t-j} \odot (1 - \mathbf{f})$ {pixels first filled from day $t - j$}
  $\mathbf{x}'_{t-1} \leftarrow \mathbf{x}'_{t-1} + \mathbf{x}_{t-j} \odot \mathbf{s}_j$
  $\mathbf{M}_{t-1}[j] \leftarrow \mathbf{s}_j$
  $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{s}_j$
end for
Future context: $\mathbf{x}'_{t+1}, \mathbf{M}_{t+1} \leftarrow$ as above, with $\mathbf{x}_{t+j}$ and $\mathbf{m}_{t+j}$
Central frame:
$\mathbf{M}_t[1] \leftarrow \mathbf{m}_t$ {observed}
$\mathbf{M}_t[2] \leftarrow 1 - \mathbf{m}_t$ {missing or land}
$\mathbf{C}_{1t} \leftarrow [\mathbf{x}'_{t-1}, \mathbf{x}_t, \mathbf{x}'_{t+1}]$
$\mathbf{C}_{2t} \leftarrow [\mathbf{M}_{t-1}, \mathbf{M}_t, \mathbf{M}_{t+1}]$

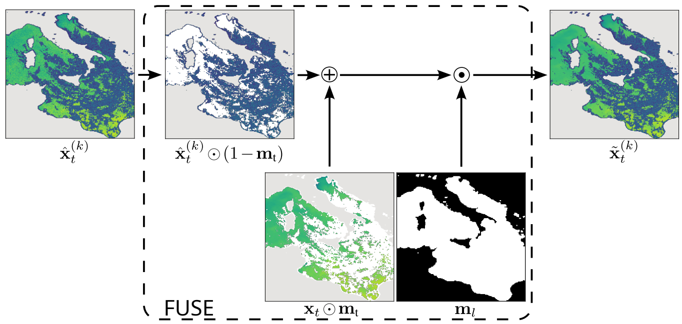

Figure 3. The FUSE operation. At each integration step k, the current estimate $\hat{\mathbf{x}}_t^{(k)}$ is merged with the original observations $\mathbf{x}_t$. This ensures that measured pixels remain preserved (non-corrupted) while the model updates the missing regions.

2.1 A flow-matching architecture

We frame the SST reconstruction as a conditional flow matching problem (Lipman et al., 2023), where the reconstruction trajectory is represented as a deterministic transport from an initial simple noise distribution toward the target data distribution. The process involves simulating a stochastic differential equation, for which the Euler integration at iteration step $k \in [0,1]$ is

$$\hat{\mathbf{x}}_t^{(k+\Delta)} = \tilde{\mathbf{x}}_t^{(k)} + \Delta \cdot \mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t), \tag{1}$$

where Δ is the integration step, $\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t)$ is the rectified flow network, with $\mathbf{C}_t$ representing the aggregated spatio-temporal context, d_t the day-of-the-year variable, and $\tilde{\mathbf{x}}_t^{(k)}$ the previous iteration reconstruction $\hat{\mathbf{x}}_t^{(k)}$ with observed values re-injected via the FUSE operation (Fig. 3), i.e.,

$$\tilde{\mathbf{x}}_t^{(k)} = \mathrm{FUSE}(\hat{\mathbf{x}}_t^{(k)}) = (\hat{\mathbf{x}}_t^{(k)} \odot (1 - \mathbf{m}_t) + \mathbf{x}_t \odot \mathbf{m}_t) \odot \mathbf{m}_l, \tag{2}$$

where $\odot$ is the Hadamard (element-wise) product. By re-injecting the clean observations at every step, the learned updates of the network $\mathbf{v}_\theta$ are restricted only to regions with missing SST values.

Figure 4 illustrates the architecture of the rectified flow network $\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t)$. The current reconstruction state $\tilde{\mathbf{x}}_t^{(k)}$ and the temporal SST context $\mathbf{C}_{1t}$ are concatenated and projected into a feature representation through a 1×1 convolution. In parallel, the temporal-origin masks $\mathbf{C}_{2t}$ are independently embedded using a separate 1×1 convolution and subsequently added to the feature tensor. The resulting representation defines the network input $\mathbf{y}_t^{(k)} \in \mathbb{R}^{128 \times W \times H}$. The backbone follows the U-Net design of (Lipman et al., 2024), where the encoder progressively increases the feature dimensionality across four resolution levels with channel widths (128, 256, 384, 512). The highest-resolution stage uses five residual blocks, while the remaining stages use three residual blocks each. Self-attention is applied only at the lowest spatial resolution.

Feature modulation is performed using FiLM conditioning (Perez et al., 2018). A global conditioning vector $\mathbf{e} \in \mathbb{R}^{512}$ is constructed by summing two embeddings: (i) Seasonal information is encoded from the day-of-year variable d_t using its sine-cosine representation $[\sin(d_t \frac{2\pi}{365.25}), \cos(d_t \frac{2\pi}{365.25})]$, mapped through a two-layer MLP; and (ii) The current flow iteration

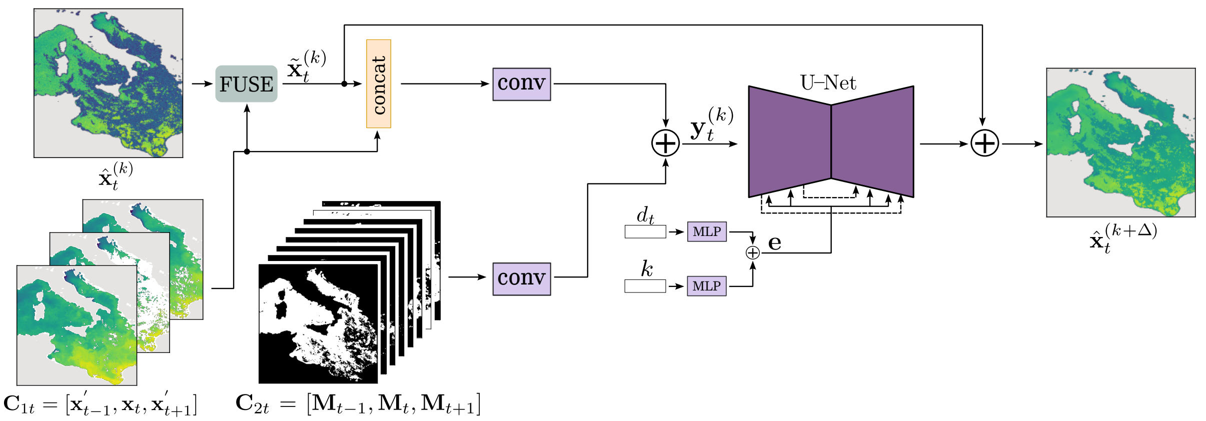

Figure 4. Overview of the DIRECT architecture. The SST field $\hat{\mathbf{x}}_t^{(k)}$ estimated at k-th iteration is accompanied by the temporal context $\mathbf{C}_{1t}$ and the time-stamp masks $\mathbf{C}_{2t}$, seasonal context d_t, and the current flow iteration value k. The FUSE rectified reconstruction $\tilde{\mathbf{x}}_t^{(k)}$ is concatenated with context frames $\mathbf{C}_{1t}$ and summed with the embedded origin masks $\mathbf{C}_{2t}$ to form the network input. Token $\mathbf{e}$ modulates the U-Net via FiLM (Perez et al., 2018). The network output updated the initial SST estimate into $\hat{\mathbf{x}}_t^{(k+\Delta)}$ by Euler integration.

variable k is encoded separately using sinusoidal embeddings and an additional two-layer MLP. The resulting conditioning token modulates feature maps throughout all residual blocks of the U-Net.

2.2 Probabilistic reconstruction

To obtain probabilistic SST estimates, the reconstruction process is repeated $N = 16$ times using different Gaussian initializations in the missing regions, producing an ensemble of reconstructions $\mathbf{S} = \{\hat{\mathbf{x}}_{t,n}\}_{n=1}^{N}$. Ideally, the ensemble spread should approximate the posterior uncertainty at each pixel. However, we observed that the ensemble variance occasionally becomes under-dispersed at isolated spatial locations. To address this effect, we model the ensemble as a mixture of Gaussians, each centered at ensemble member with a small constant variance σ_0^2 across all pixels. The final reconstruction is thus obtained by moment-matching the mixture with a single Gaussian, i.e.,

$$\boldsymbol{\mu}_t = < \mathbf{S} > ; \; \boldsymbol{\sigma}_t^2 = \mathrm{var}(\mathbf{S}) + \sigma_0^2, \tag{3}$$

where the constant σ_0^2 is estimated on the validation set (explained in the next section).

2.3 Training procedure

To enable training with incomplete data, DIRECT is trained by a modified self-supervised flow-matching objective. The training samples are created by sampling cloud masks $\mathbf{m}_m$ from other days and applying them to the central day, yielding $\mathbf{x}''_t = \mathbf{x}_t \odot \mathbf{m}_m$ (Fig. 5). Thus, a ground truth direction vector $\mathbf{u}_t = \mathbf{x}_t - \tilde{\mathbf{x}}_t''^{(0)}$ is constructed from the original observed field ($\mathbf{x}_t$) and the simulated observed field with missing values initialized by noise ($\tilde{\mathbf{x}}_t''^{(0)}$). The flow iteration variable k is sampled

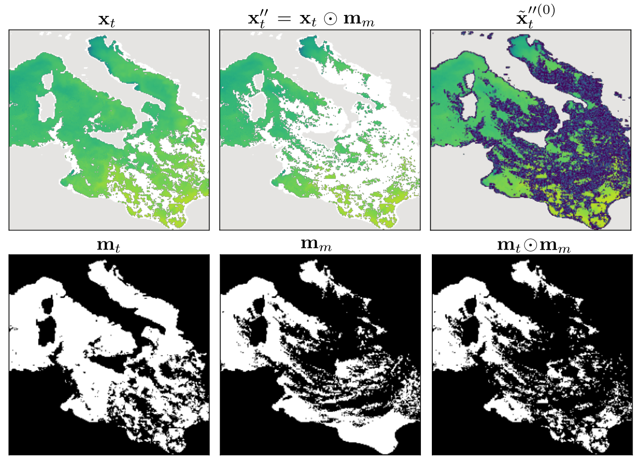

Figure 5. First row: ground truth incomplete central day observation ($\mathbf{x}_t$), simulated central day observation ($\mathbf{x}''_t$), noised initialized simulated central day observation ($\tilde{\mathbf{x}}''^{(0)}_t$). Second row: mask denoting valid observations in ground truth observation ($\mathbf{m}_t$), sampled mask from a different day ($\mathbf{m}_m$), mask denoting observed regions in the simulated central day observation ($\mathbf{m}_t \odot \mathbf{m}_m$).

at uniform from interval $[0,1]$, leading to the following loss at field indexed by time-step t:

$$\mathcal{L}_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\left[\left(\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)}|\mathbf{C}_t,k,d_t)_{(i)} - \mathbf{u}_{t,i}\right)^2 \mathbf{m}_{t(i)}\right], \tag{4}$$

where $N_t = |\mathbf{m}_t|$ is the number of observed pixels in the ground truth SST field. Note that because complete cloud-free data is unavailable, we cannot supervise in the originally clouded regions. As such, this simulation-based approach allows the model to learn the underlying spatial correlations from the available data.

After training the model, the training set (or validation if available) can be reconstructed and the regularization variance σ_0^2 estimated as follows. Following (Zupančič Muc et al., 2025) we define per-pixel scaled reconstruction error as

$$\boldsymbol{\epsilon} = \frac{\mathbf{x}_t - \boldsymbol{\mu}_t}{\boldsymbol{\sigma}_t}. \tag{5}$$

For a well-calibrated estimator, this error should follow a unit-variance, zero-centered Gaussian, i.e., $\mu_\epsilon \to 0$ and $\sigma_\epsilon \to 1$, where μ_ϵ and σ_ϵ are the empirical mean and standard deviation of $\boldsymbol{\epsilon}$, calculated over the training (or validation) set. The regularization variance σ_0^2 can thus be estimated by minimizing the KL divergence (Kullback and Leibler, 1951) between the zero-mean, unit Gaussian and the empirical Gaussian obtained from (5), which yields the following loss

$$\mathcal{L}_{\sigma_0^2} = \mu_\epsilon^2 + \sigma_\epsilon^2 - \log\sigma_\epsilon^2. \tag{6}$$

3 Training, Validation and Testing Data

To assess DIRECT, we adopt three datasets corresponding to Mediterranean, Adriatic and Atlantic regions. These regions were chosen because they exhibit very diverse oceanographic regimes. Mediterranean basin contains several characteristic subbasins, including regions of freshwater influence, frontal systems and regions with submesoscale and mesoscale eddy activity. Adriatic has a shallow shelf in the north, highly influenced by local freshwater discharges, while its southern part is deeper and exposed to Atlantic and Levantine medium waters entering through Otranto strait from the Ionian sea. Finally, the Atlantic region is essentially an open ocean, marked by intense cyclonic activity and cloudiness. All datasets are based on level-3 (L3) multi-sensor SST products and are provided as daily fields remapped onto uniform regular grids. For each dataset, we employ chronological splits to prevent temporal leakage: first 90 % of the samples are used for training, further 5 % are used for validation (to monitor model performance and guide design choices), and final 5 % are used for testing.

1. The Mediterranean SST_MED_SST_L3S_NRT_OBSERVATIONS_010_012 (E.U. Copernicus Marine Service Information, 2024c) dataset consists of daily SST observations over the Central Mediterranean from January 1, 2008, to December 31, 2021, remapped to a $0.0625° \times 0.0625°$ grid.

2. The Adriatic SST_MED_PHY_L3S_MY_010_042 (E.U. Copernicus Marine Service Information, 2024a) dataset contains daily SST fields over the Adriatic Sea from August 25, 1981, to December 31, 2022, remapped to a $0.05° \times 0.05°$ grid.

3. The Atlantic SST_ATL_PHY_L3S_MY_010_038 (E.U. Copernicus Marine Service Information, 2024b) dataset spans an open-ocean region in the North Atlantic from January 1, 1982, to January 1, 2022, remapped to a $0.05° \times 0.05°$ grid.

Following Zupančič Muc et al. (2025), we apply a coverage threshold to the central observation in each input triplet to filter out samples with excessive cloud cover. A sample is discarded if more than a dataset-specific percentage of the central frame is missing: 100 % for the Mediterranean, 60 % for the Adriatic, and 75 % for the Atlantic dataset. After filtering, we obtain 5114 valid samples for the Mediterranean, 7800 for the Adriatic, and 3454 for the Atlantic dataset.

4 Results

4.1 Performance measures

We report the Root Mean Squared Error (RMSE) calculated over three specific regions: the region with observed data ($\mathrm{RMSE}_{\mathrm{obs}}$), the region in which we occluded the data ($\mathrm{RMSE}_{\mathrm{mis}}$), and the two regions combined for overall evaluation ($\mathrm{RMSE}_{\mathrm{all}}$). While $\mathrm{RMSE}_{\mathrm{obs}}$ is included for completeness, it is not particularly informative: due to the FUSE rectification step (Eq. (2)), observed pixels are replaced with their measured values at every flow integration step, leading to near-zero error. Nevertheless, note that many competing methods also reconstruct this part of the data, necessarily increasing the errors also in the *observed* regions. The metric of primary interest is therefore $\mathrm{RMSE}_{\mathrm{mis}}$, which reflects the quality of reconstruction in the occluded regions.

4.2 Comparison with state-of-the-art

DIRECT is evaluated against four recent SST reconstruction approaches: DINCAE2 (Barth et al., 2022), CARE-SST (Choo et al., 2025), MAESSTRO (Goh et al., 2024), and CRITER (Zupančič Muc et al., 2025). We follow the evaluation procedure introduced by Zupančič Muc et al. (2025), in which each test sample is evaluated under 10 independently sampled cloud-mask realizations applied to the central frame. Final performance scores are obtained by averaging the reconstruction errors across all repetitions and all samples in the test dataset.

Results in Table 1 show that across all three datasets, DIRECT achieves the lowest reconstruction error among all considered methods. Relative to DINCAE2, DIRECT reduces $RMSE_{mis}$ by 52 % on the Adriatic dataset, 27 % on the Mediterranean, and 7 % on the Atlantic dataset. Improvements over MAESSTRO are even greater, ranging from 41 % on the Atlantic to more than 62 % on both the Mediterranean and Adriatic datasets. Notably, because CARE-SST was originally designed for fully-observed ground truth, we adapted its training objective to our self-supervised formulation (Sec. 2.3) to enable learning from incomplete data. Even with this adaptation, DIRECT achieves $RMSE_{mis}$ by 55 % lower on the Mediterranean, 51 % on the Adriatic, and 36 % on the Atlantic. Compared to CRITER, the strongest published baseline, DIRECT further reduces reconstruction error in occluded regions by 14 % on the Adriatic, 8 % on the Mediterranean, and 6 % on the Atlantic. The quantitative improvements are also reflected in the spatial consistency of the reconstructed SST fields. Figures 6, 7, and 8 show DIRECT and CRITER reconstruction examples from the Mediterranean, Adriatic, and Atlantic datasets, respectively. Across the Mediterranean and Adriatic examples, DIRECT better recovers the underlying SST structure under a wide variety of cloud shapes and coverage levels. For the Atlantic domain, where cloud cover is typically more extensive and ocean variability more energetic, CRITER reconstructions often exhibit visible patch artifacts, a byproduct of its patch-based processing. DIRECT avoids such artifacts due to its fully convolutional diffusion backbone and produces more spatially consistent fields even under extreme occlusions. For subsequent experiments, we retain CRITER as the sole baseline, since it consistently outperforms DINCAE2 and MAESSTRO.

4.2.1 Comparison with operational L4 products

We compare DIRECT against operational gap-free Level-4 (L4) SST products, based on variational data assimilation and optimal interpolation, distributed by the Copernicus Marine Service: the reprocessed OSTIA analysis (E.U. Copernicus Marine Service Information (CMEMS), 2026b), the ESA SST CCI and C3S reprocessed analyses (E.U. Copernicus Marine Service Information (CMEMS), 2026a). Each product was co-located with SST from our three domain datasets and evaluated as a candidate reconstruction against the same L3 reference used throughout. This comparison should be read as a measure of agreement with our L3 reference, not of reconstruction accuracy. The L4 products also ingest in situ data and data from the same sensors as our L3 source, meaning the pixels we synthetically withhold were in general available to the L4 products. Their error over these pixels is therefore not an out-of-sample gap-filling error as it is for DIRECT. The L4 products additionally carry a systematic offset relative to our L3 reference (Tab. 2, bias column); to prevent this offset from dominating the comparison, we estimate a per-day mean bias over the observed region, subtract it from each L4 field, and report the resulting centered RMSE (CRMSE).

Table 2 shows that, even after this bias removal, DIRECT remains closer to the L3 reference than any of the operational products in all three regions. Relative to the best-performing L4 product in each region, $CRMSE_{mis}$ is lower by 44 % on the Mediterranean, 37 % on the Adriatic, and 8 % on the Atlantic, indicating that the remaining gap reflects a difference in spatial

Table 1. Comparison of DIRECT to current state-of-the-art methods. All of the reported reconstruction errors are in °C, where the two numbers in parentheses correspond to the 10 % and 90 % percentiles.

Dataset	Model	$\mathrm{RMSE}_{\mathrm{all}}$ [°C]	$\mathrm{RMSE}_{\mathrm{mis}}$ [°C]	$\mathrm{RMSE}_{\mathrm{obs}}$ [°C]
Mediterranean	MAESSTRO (Goh et al., 2024)	0.487 (0.320, 0.657)	0.607 (0.394, 0.856)	0.434 (0.299, 0.564)
	CARE-SST (Choo et al., 2025)	0.268 (0.064, 0.500)	0.518 (0.316, 0.716)	0.036 (0.034, 0.038)
	DINCAE2 (Barth et al., 2022)	0.209 (0.140, 0.300)	0.319 (0.226, 0.418)	0.148 (0.112, 0.184)
	CRITER (Zupančič Muc et al., 2025)	0.127 (0.037, 0.235)	0.255 (0.168, 0.352)	0.017 (0.013, 0.021)
	DIRECT (ours)	**0.113 (0.027, 0.211)**	**0.234 (0.150, 0.320)**	**0.000 (0.000, 0.001)**
Adriatic	MAESSTRO (Goh et al., 2024)	0.456 (0.296, 0.635)	0.583 (0.362, 0.844)	0.392 (0.261, 0.539)
	CARE-SST (Choo et al., 2025)	0.243 (0.070, 0.404)	0.428 (0.241, 0.620)	0.036 (0.034, 0.039)
	DINCAE2 (Barth et al., 2022)	0.270 (0.111, 0.522)	0.433 (0.203, 0.769)	0.106 (0.087, 0.129)
	CRITER (Zupančič Muc et al., 2025)	0.130 (0.045, 0.222)	0.243 (0.140, 0.358)	0.021 (0.014, 0.030)
	DIRECT (ours)	**0.113 (0.029, 0.200)**	**0.208 (0.116, 0.312)**	**0.001 (0.001, 0.002)**
Atlantic	MAESSTRO (Goh et al., 2024)	0.802 (0.508, 1.239)	0.832 (0.514, 1.301)	0.764 (0.479, 1.137)
	CARE-SST (Choo et al., 2025)	0.575 (0.336, 0.818)	0.767 (0.544, 1.009)	0.029 (0.026, 0.032)
	DINCAE2 (Barth et al., 2022)	0.444 (0.332, 0.581)	0.525 (0.396, 0.692)	0.302 (0.236, 0.364)
	CRITER (Zupančič Muc et al., 2025)	0.391 (0.249, 0.542)	0.518 (0.386, 0.692)	0.036 (0.019, 0.046)
	DIRECT (ours)	**0.363 (0.236, 0.493)**	**0.489 (0.370, 0.629)**	**0.001 (0.000, 0.002)**

structure rather than a mean offset. Note that this does not imply that DIRECT is correspondingly more accurate than objective analysis as a technique; it reflects that DIRECT is trained to reconstruct this specific high-resolution L3 product, whereas the L4 analyses target foundation SST and cross-sensor, temporally consistent fields rather than fidelity to any one regional L3 source. A fair skill comparison would evaluate both against independent in-situ observations such as Argo profiles, which neither assimilates, a direction we leave to future work.

4.2.2 Comparison under varying cloud coverage

To assess the robustness of DIRECT to varying levels of missing data, we evaluate reconstruction performance across three cloud-coverage regimes, defined by the fraction of missing pixels in the central observation $\mathbf{x}_t$ relative to all sea pixels. We define low coverage in the interval [0 %, 60 %], moderate coverage in the interval (60 %, 75 %], and high coverage in the interval (75 %, 100 %]. The test sets span a wide range of coverage: 8.3–99 % for the Mediterranean, 3.3–95 % for the Adriatic, and 39–99 % for the Atlantic.

Figure 9 shows the reconstruction errors ($\mathrm{RMSE}_{\mathrm{all}}$, $\mathrm{RMSE}_{\mathrm{mis}}$, $\mathrm{RMSE}_{\mathrm{obs}}$) for DIRECT and CRITER across the three regimes, where, as before, $\mathrm{RMSE}_{\mathrm{obs}}$ is reported for completeness, but provides little insight due to the injection of ground truth values at

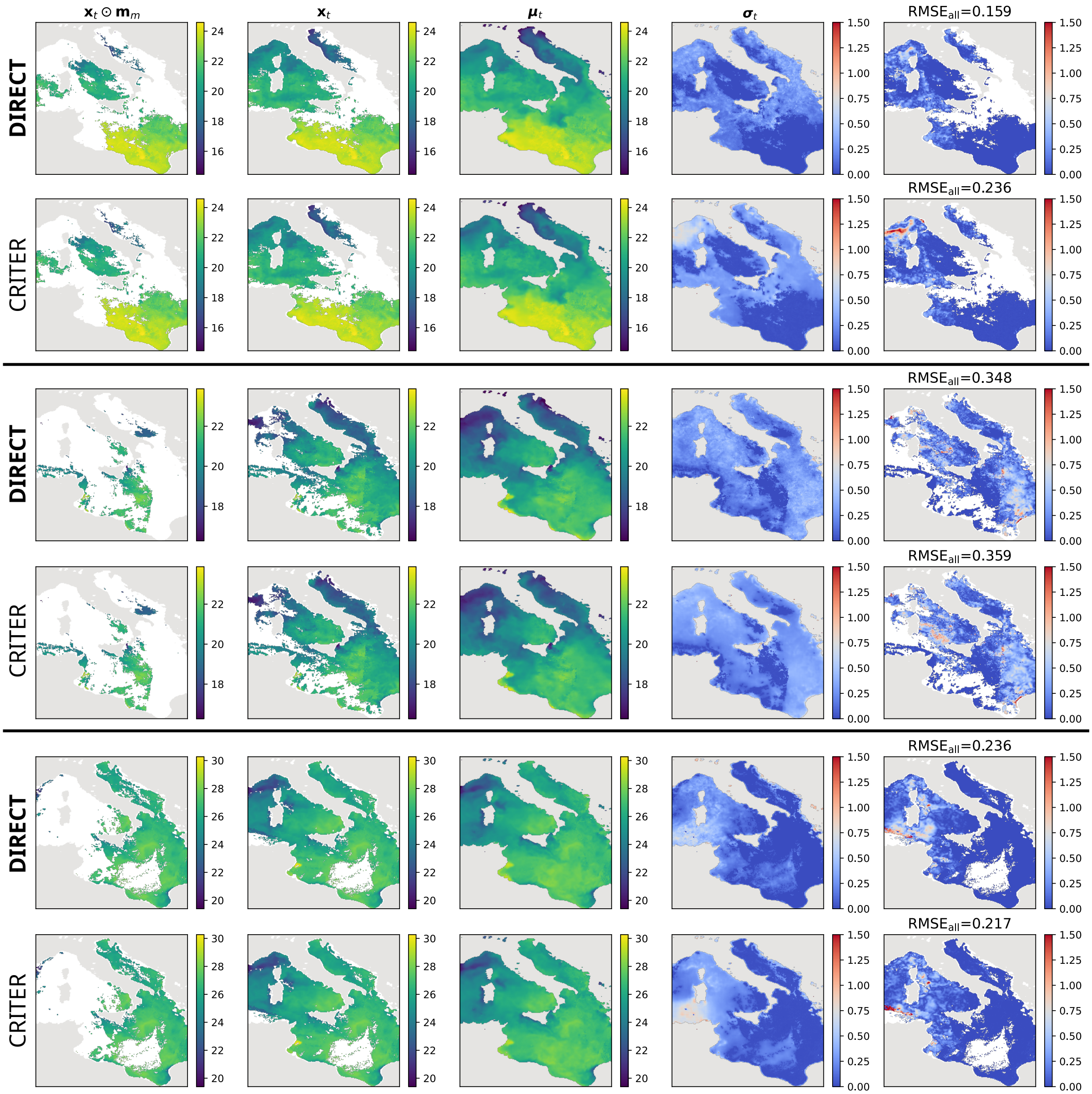

Figure 6. Comparison of DIRECT and CRITER reconstructions on the Mediterranean region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE$_{\text{all}}$). All values are in $^{\circ}$C.

observed locations. Across all cloud-coverage levels and for all three regions, DIRECT consistently outperforms CRITER. On the Mediterranean, RMSE$_{\text{mis}}$ is reduced by 8 %, 6 %, and 10 % in the low-, medium-, and high-coverage regimes, respectively. Improvements are even greater on the Adriatic, with reductions of 16 %, 12 %, and 9 %. On the Atlantic, where occlusions are

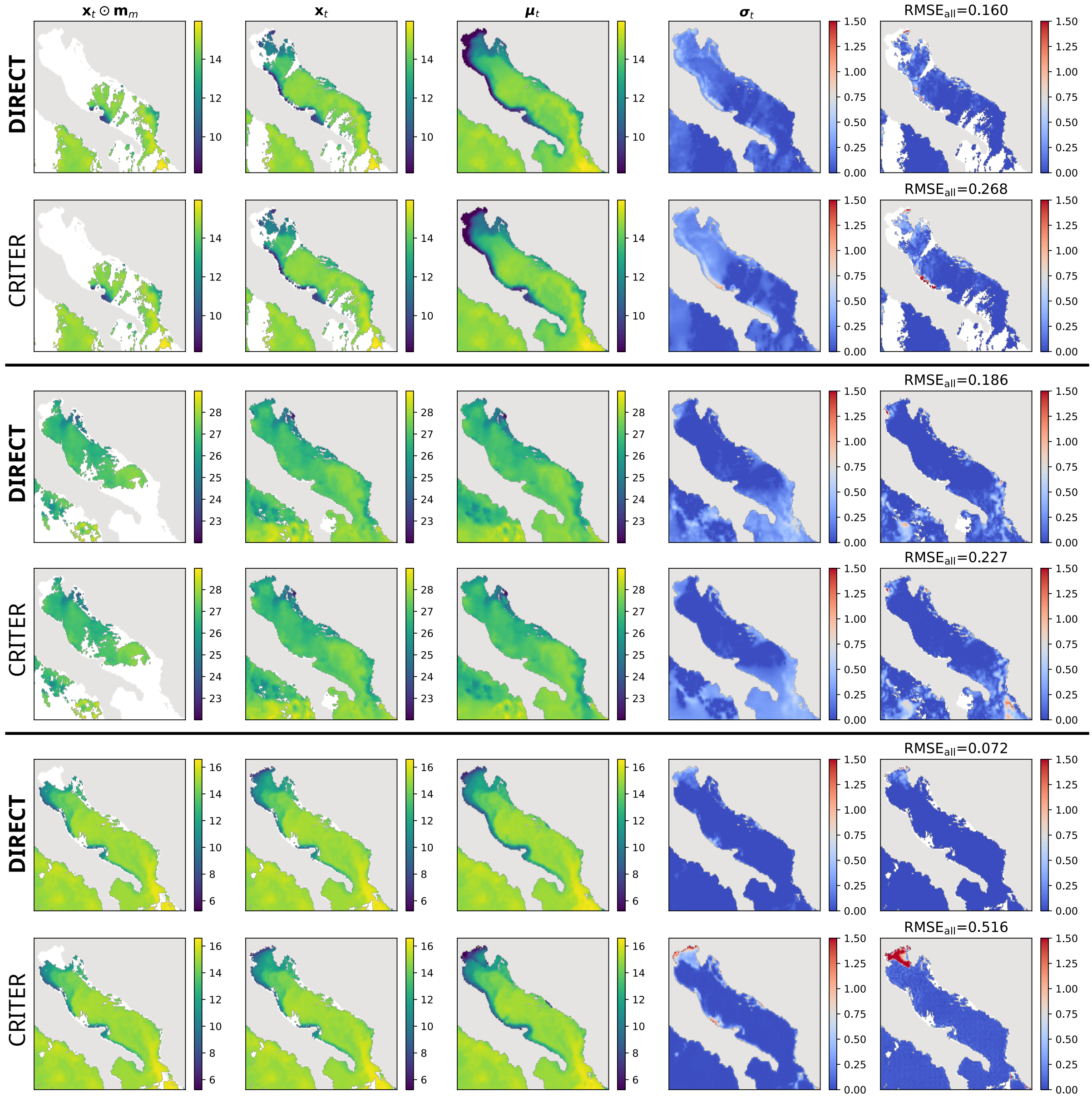

Figure 7. Comparison of DIRECT and CRITER reconstructions on the Adriatic region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE$_{\text{all}}$). All values are in °C.

more extreme and the problem is inherently more challenging, smaller yet meaningful improvements of 2 %, 4 %, and 6 % are achieved.

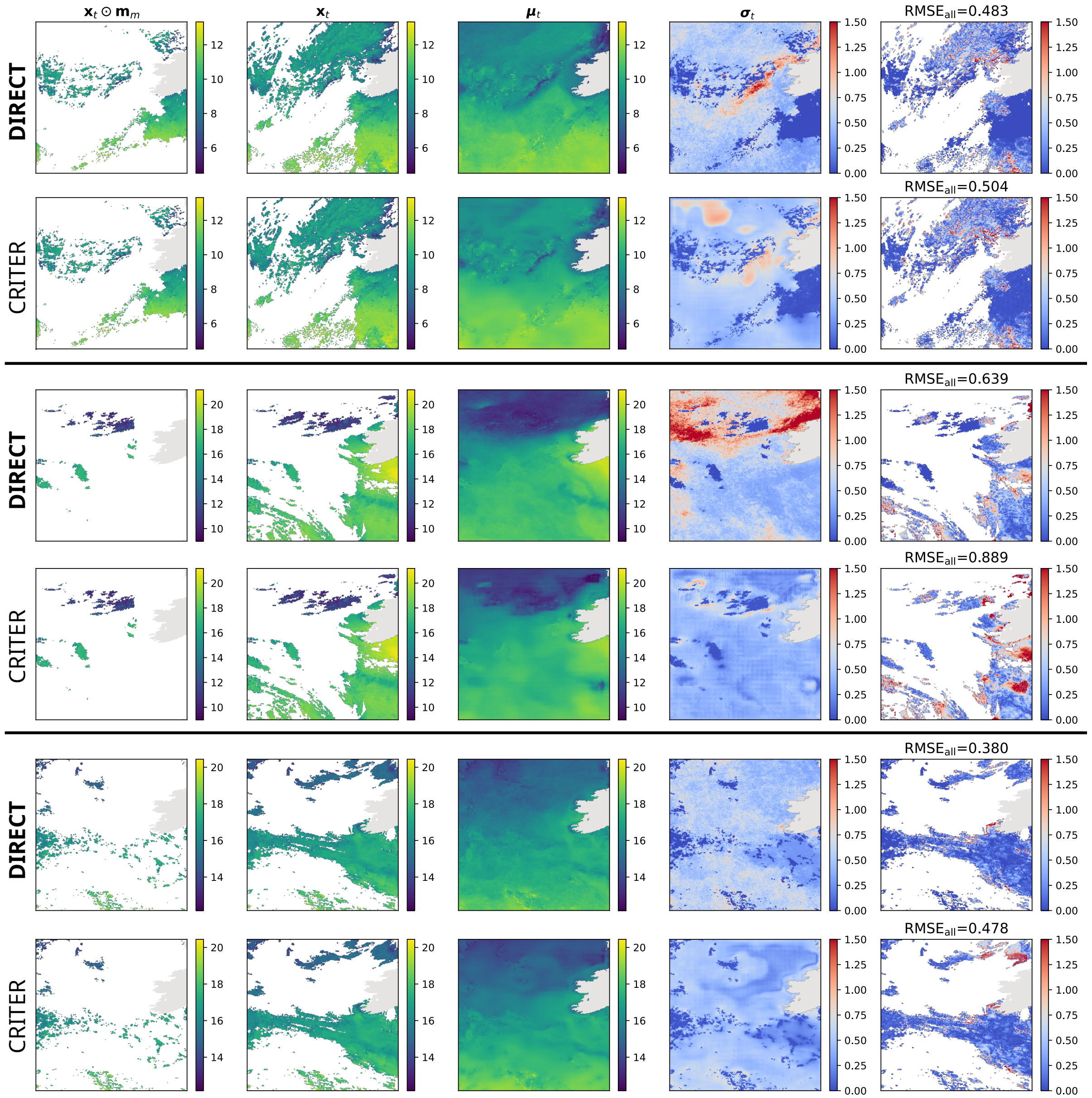

Figure 8. Comparison of DIRECT and CRITER reconstructions on the Atlantic region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE$_{\text{all}}$). All values are in °C.

Table 2. Comparison of DIRECT against operational gap-free L4 products, evaluated against the same L3 reference. Note that $\mathrm{CRMSE}_{\mathrm{mis}}$ is an error against withheld, independent data only for DIRECT; the L4 analyses ingest the same sensors as our L3 reference, so these pixels were in general available to them. Bias is calculated over the observed region. Errors in °C

Dataset	Model / Product	$\mathrm{CRMSE}_{\mathrm{all}}$ [°C]	$\mathrm{CRMSE}_{\mathrm{mis}}$ [°C]	$\mathrm{CRMSE}_{\mathrm{obs}}$ [°C]	Bias [°C]
Mediterranean	DIRECT (ours)	**0.113**	**0.234**	**0.000**	0.000
	ESA CCI	0.401	0.421	0.393	0.239
	C3S	0.399	0.420	0.392	0.283
	OSTIA	0.408	0.429	0.400	0.241
Adriatic	DIRECT (ours)	**0.113**	**0.208**	**0.001**	0.000
	ESA CCI	0.301	0.329	0.290	0.186
	C3S	0.324	0.355	0.313	0.175
	OSTIA	0.348	0.378	0.336	0.119
Atlantic	DIRECT (ours)	**0.363**	**0.489**	**0.001**	0.000
	ESA CCI	0.533	0.542	0.520	0.235
	C3S	0.522	0.531	0.510	0.275
	OSTIA	0.540	0.550	0.528	0.356

4.2.3 Uncertainty estimation and bias analysis

We evaluate the reliability of ensemble-based uncertainty estimates over the withheld regions using two complementary measures: the scaled error $\boldsymbol{\epsilon}$ of Eq. (5) and the empirical coverage of the nominal 50 %, 80 %, 90 %, and 95 % central prediction intervals, where a withheld pixel is covered at level p when $|\,\boldsymbol{\epsilon}\,| \leq z_p$ for the standard-normal quantile z_p. Results for the raw ensemble and after correction are given in Tables 3 and 4. Without the correction ($\mathrm{DIRECT}_{\mathrm{w.o.}\sigma_0}$), the ensemble is under-dispersed, with $\sigma_\epsilon > 1.7$ and coverage below nominal at every level and in all three regions. The two measures together indicate two distinct effects. Coverage is depressed across the full range of levels, including the central 50 % interval, which reflects a broad under-dispersion of the ensemble. At the same time, the large σ_ϵ points to pronounced ensemble collapse at a sparse set of pixels, where the samples become nearly identical and the predicted variance is far too small. The learned regularization constant σ_0 (0.135 for Mediterranean, 0.101 for Adriatic, 0.298 for Atlantic), added in Eq. (3), addresses both effects. It raises the overall spread, correcting the global under-dispersion, and places a floor on the variance at collapsed pixels. With the correction, σ_ϵ falls close to one and coverage approaches nominal at the 80 %, 90 %, and 95 % levels across all datasets. As a single global offset applied uniformly across pixels and probability levels, σ_0 does not achieve exact calibration everywhere: the corrected 50 % intervals slightly over-cover (about 59-61 %) and the 95 % intervals slightly under-cover (about 93-94 %), since one constant cannot adapt to spatially varying error regimes or to the tails of the distribution. Overall, σ_0 accounts for 34 %, 25 %, and 36 % of the total variance across the Mediterranean, Adriatic, and Atlantic datasets, respectively.

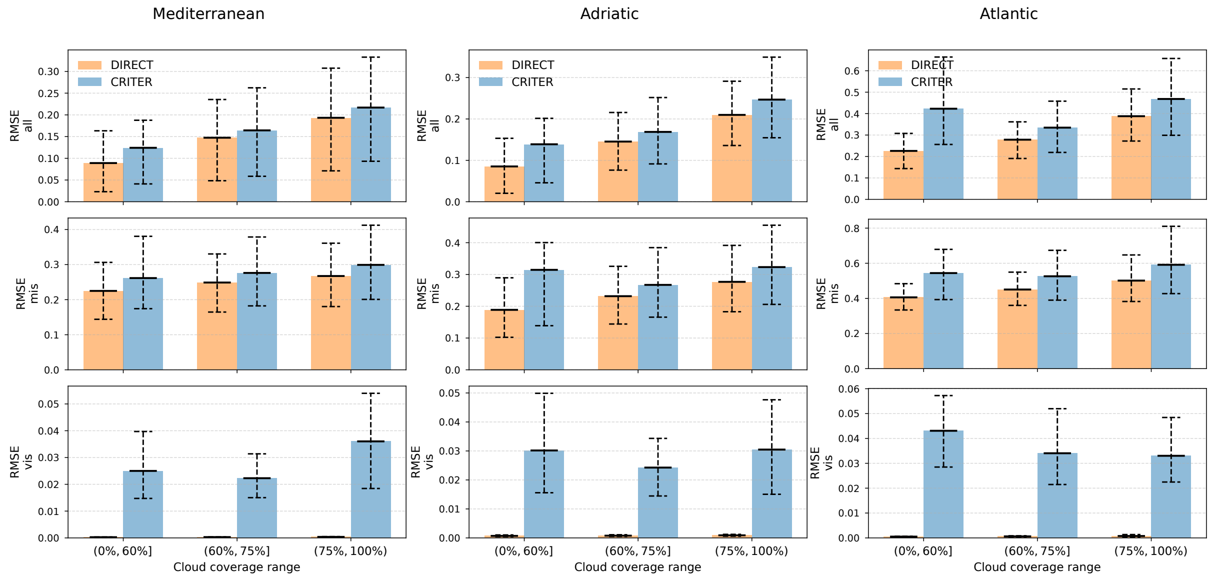

Figure 9. Reconstruction error comparison of DIRECT and CRITER under varying levels of cloud coverage (low, moderate, and high) on the Mediterranean, Adriatic, and Atlantic datasets. The error bars indicate the 10 % percentile, mean, and the 90 % percentile of the error.

Table 3. Uncertainty estimation analysis of DINCAE2, CRITER, and DIRECT. Mean standardized error (μ_ϵ), standard deviation (σ_ϵ), and bias are reported for each dataset.

Dataset	Model	μ_ϵ	σ_ϵ	Bias
Mediterranean	DINCAE2	-0.060	0.334	-0.060
	CRITER	-0.022	1.116	-0.007
	DIRECT$_{\mathrm{w.o.}\sigma_0}$	-0.036	1.728	**-0.003**
	DIRECT	**-0.018**	**1.059**	**-0.003**
Adriatic	DINCAE2	0.198	**0.996**	0.128
	CRITER	0.041	1.082	0.007
	DIRECT$_{\mathrm{w.o.}\sigma_0}$	0.029	1.733	**-0.004**
	DIRECT	**0.018**	1.018	**-0.004**
Atlantic	DINCAE2	**-0.017**	0.801	**-0.006**
	CRITER	0.118	1.156	0.047
	DIRECT$_{\mathrm{w.o.}\sigma_0}$	-0.080	1.792	-0.008
	DIRECT	-0.035	**1.087**	-0.008

Table 4. Empirical coverage of nominal central prediction intervals, before (DIRECT$_{\mathrm{w.o.}\sigma_0}$) and after (DIRECT) the σ_0 correction.

Dataset	Nominal	DIRECT$_{\mathrm{w.o.}\sigma_0}$	DIRECT
Mediterranean	50%	36.8%	58.7%
	80%	62.1%	82.0%
	90%	72.6%	89.1%
	95%	79.4%	93.0%
Adriatic	50%	36.5%	61.2%
	80%	61.8%	83.3%
	90%	72.5%	90.1%
	95%	79.5%	93.7%
Atlantic	50%	39.4%	58.6%
	80%	65.2%	83.2%
	90%	75.2%	89.8%
	95%	81.3%	93.1%

4.2.4 Power spectrum density analysis

To evaluate how well DIRECT preserves spatial variability across scales, we perform a power spectral density (PSD) analysis (Stoica et al., 2005) on a region of interest (ROI) in the central Mediterranean and compare it with CRITER (Zupančič Muc et al., 2025). Time frames for which the ROI is fully observed are first identified, after which cloud masks are sampled, obscuring between 51 % and 100 % of the ROI. Both models reconstruct the full SST field from these masked inputs. For each reconstruction, we first remove the large-scale spatial background by subtracting a least-squares linear plane fitted over the ROI, which suppresses the anisotropic low-wavenumber trend. We then compute the 2D gradient magnitude $||\nabla \boldsymbol{\mu}_t^{\mathrm{ROI}}||_2$ and the isotropic PSD using a FFT with a Blackman–Harris window (Harris, 1978).

Figures 10d and 11d show that both models reproduce the large-scale spectral structure of the SST field and follow the ground-truth spectrum closely from the largest scales down to intermediate wavenumbers. At higher wavenumbers both models depart from the ground truth and settle at similar energy levels. For CRITER, however, this small-scale energy is largely non-physical: the gradient-magnitude fields (Figs. 10c and 11c) reveal grid-like block artifacts, a byproduct of patch-based processing. In contrast, DIRECT's energy profile reflects coherent, albeit slightly smoothed, oceanic structures. The same gradient-magnitude fields show no discontinuity tracing the boundary between observed and reconstructed regions (Fig 11c), indicating that the FUSE re-injection does not introduce a visible seam despite anchoring the output to measured pixels at every step. DIRECT therefore provides a more physically faithful representation of the continuum, avoiding both the spurious grid-like artefacts that inflate the PSD of patch-based baselines and the boundary artifacts that might be expected from observation re-injection.

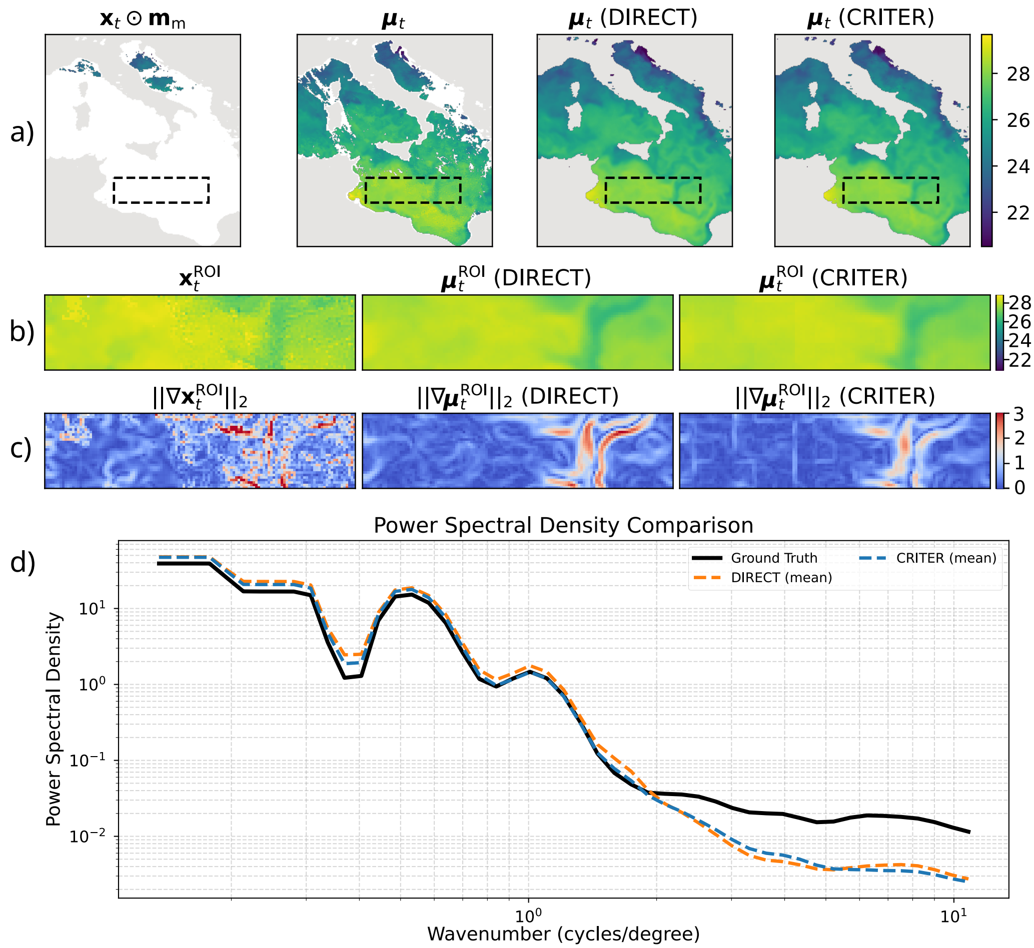

Figure 10. Power spectral density (PSD) analysis for a selected region of interest (ROI). a) masked SST input, ground truth field, and reconstructions from DIRECT and CRITER. b) zoomed views of the ROI. c) Gradient magnitudes $||\nabla \boldsymbol{\mu}_t^{\mathrm{ROI}}||_2$ highlighting structural differences. d) Isotropic PSD computed from the plane-detrended ROI fields using an FFT with a Blackman–Harris window. The displayed curves represent the average spectral density across 10 independent mask realizations.

4.2.5 Analysis of spatial scale correlation

To assess the preservation of spatial structure and mesoscale variability beyond pixel-wise error, we analyze the spatial correlation properties using empirical semivariograms (Matheron, 1963). Semivariance is computed as

$$\gamma(d) = \frac{1}{2N(d)} \sum_{i,j \in \mathcal{P}(d)} (p_i - p_j)^2, \tag{7}$$

where $\mathcal{P}(d)$ denotes the set of $N(d)$ pixel pairs separated by distance d, and p_i is the observation value of $\mathbf{x}_t$ at i-th pixel. For statistical robustness, the analysis is averaged over 10 independent mask realizations sampled from the Mediterranean dataset.

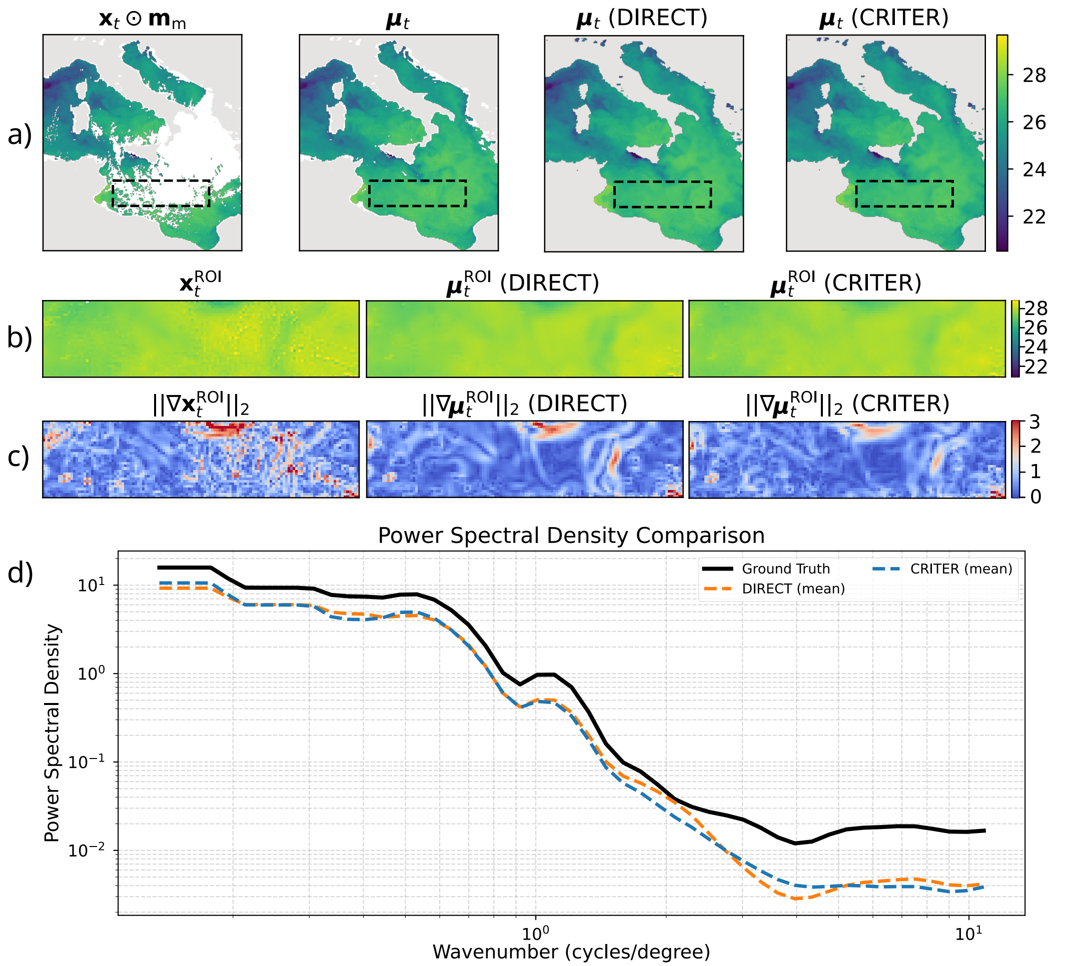

Figure 11. Power spectral density (PSD) analysis for a selected region of interest (ROI). a) masked SST input, ground truth field, and reconstructions from DIRECT and CRITER. b) zoomed views of the ROI. c) Gradient magnitudes $||\nabla \boldsymbol{\mu}_t^{\mathrm{ROI}}||_2$ highlighting structural differences. d) Isotropic PSD computed from the plane-detrended ROI fields using an FFT with a Blackman–Harris window. The displayed curves represent the average spectral density across 10 independent mask realizations.

As shown in Figs. 12 and 13, both models effectively capture local correlations (short lags). At intermediate to larger spatial lags, however, clearer differences appear. In the hidden region variograms, DIRECT consistently reproduces the growth of semivariance with distance more faithfully, while CRITER exhibits less stable long-range behavior, suggesting a struggle to maintain spatial coherence over large obscured gaps. When considering the full valid SST domain, the performance gap narrows significantly. In this setting, both DIRECT and CRITER follow the ground truth semivariance almost perfectly, with virtually no observable difference between the two models. This convergence is expected, as the calculation includes observed

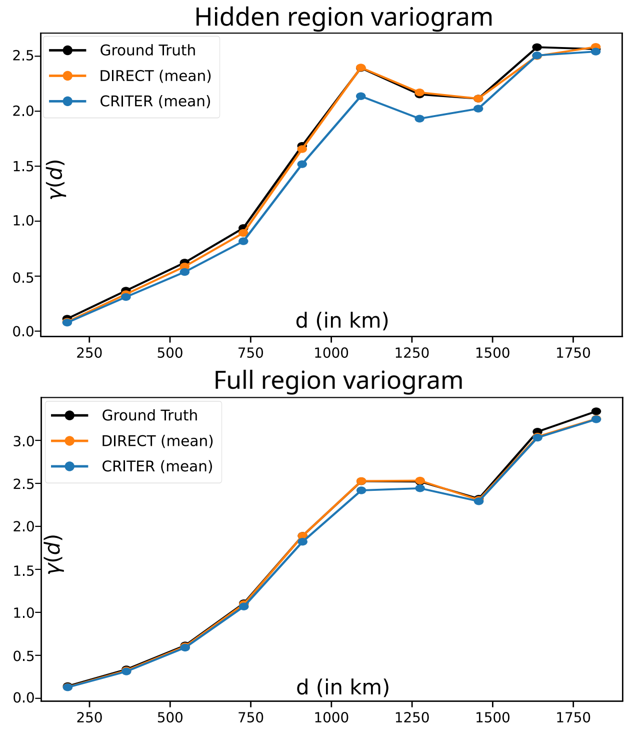

Figure 12. Visualization of the spatial scales and correlation properties analysis. Top plot: empirical semivariograms computed exclusively over hidden regions. Bottom plot: semivariograms computed over the full SST domain. Each curve represents the mean semivariance across 10 different mask realizations.

pixels. The results highlight that while both models are highly reliable when observations are present, DIRECT provides a more physically consistent spatial texture when reconstructing larger completely unobserved areas.

4.2.6 Continuous Ranked Probability Score (CRPS)

The Continuous Ranked Probability Score (CRPS) (Matheson and Winkler, 1976) was evaluated over all three regions, and results are shown in Table 5. This provides a measure for evaluating probabilistic predictions by comparing the entire forecast distribution against the observed value. Unlike traditional deterministic error metrics, which assess only the mean reconstruction, the CRPS quantifies accuracy, statistical reliability, and sharpness of the predictive distribution. Rather than simply assessing whether DIRECT obtains a plausible distribution, the CRPS allows us to qualify if the produced ensemble actually explores the inherent uncertainty in the reconstruction and if probabilities derived from the ensemble are reliable. For DIRECT, the CRPS is computed empirically from the reconstruction ensemble. As CRITER does not produce an ensemble but instead predicts a per-pixel mean and variance, we compute its CRPS by drawing samples from the corresponding per-pixel Gaussian. Across all three regions, DIRECT achieves lower CRPS than CRITER, indicating that its ensemble provides a sharper and

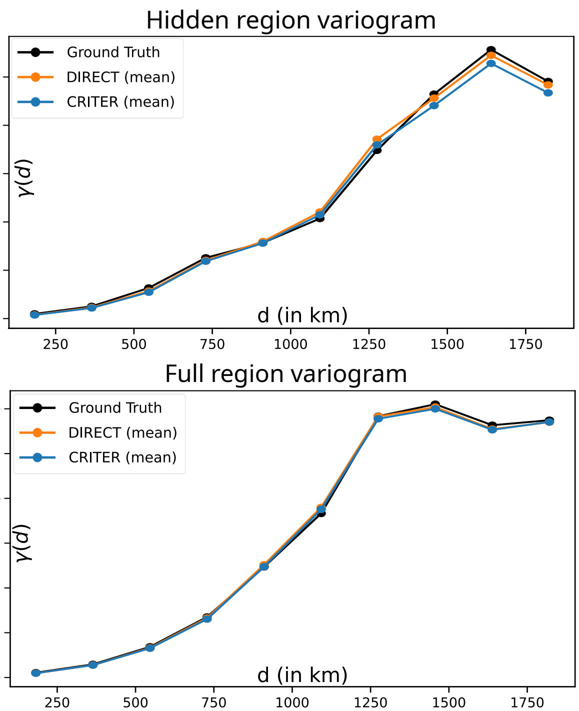

Figure 13. Visualization of the spatial scales and correlation properties analysis. Top plot: empirical semivariograms computed exclusively over hidden regions. Bottom plot: semivariograms computed over the full SST domain. Each curve represents the mean semivariance across 10 different mask realizations.

Table 5. Continuous Ranked Probability Score of CRITER and DIRECT over the whole region, the missing region, and the observed region. The values are in °C. Med. denotes the Mediterranean dataset.

Dataset	Model	$CRPS_{all}$	$CRPS_{mis}$	$CRPS_{obs}$
Med.	CRITER	0.052	0.172	0.014
	DIRECT	**0.034**	**0.116**	**0.000**
Adriatic	CRITER	0.049	0.132	0.012
	DIRECT	**0.033**	**0.097**	**0.000**
Atlantic	CRITER	0.199	0.316	0.037
	DIRECT	**0.145**	**0.255**	**0.001**

better-calibrated predictive distribution. As with the pointwise reconstruction errors, CRPS over the missing pixels is highest in the Atlantic region, reflecting the more extensive cloud cover and more energetic ocean variability.

Table 6. Performance of DIRECT variants. All of the reported reconstruction errors are in °C.

Variant	$\mathrm{RMSE}_{\mathrm{all}}$	$\mathrm{RMSE}_{\mathrm{mis}}$	$\mathrm{RMSE}_{\mathrm{obs}}$
$\mathrm{DIRECT}_{\overline{\mathrm{OBS}}}$	0.115	0.237	0.011
$\mathrm{DIRECT}_{\overline{\mathrm{ML}}}$	0.115	0.239	0.000
$\mathrm{DIRECT}_{\overline{\mathrm{FUSE}}}$	0.222	0.295	0.181
DIRECT	**0.113**	**0.234**	**0.000**

4.3 Ablation study

Ablation studies were performed on the Mediterranean dataset, unless stated otherwise, to expose the influences of individual parts.

4.3.1 Importance of input rectification (FUSE)

We study the effect of the FUSE procedure (Eq. 2) by comparing DIRECT with three ablated variants: one without injecting observed SST values ($\mathrm{DIRECT}_{\overline{\mathrm{OBS}}}$), one without zeroing land pixels ($\mathrm{DIRECT}_{\overline{\mathrm{ML}}}$), and one without any rectification ($\mathrm{DIRECT}_{\overline{\mathrm{FUSE}}}$). Results in Table 6 show that completely disabling FUSE causes a large drop in reconstruction accuracy, with $\mathrm{RMSE}_{\mathrm{mis}}$ increased by more than 20 %. Restoring only the injection of SST values ($\mathrm{DIRECT}_{\overline{\mathrm{ML}}}$) or only the masking of land pixels ($\mathrm{DIRECT}_{\overline{\mathrm{OBS}}}$) recovers much of this loss (only a 2 % increase for both), indicating that both corrections help anchor the generative process and prevent errors from accumulating across integration steps. The best results are obtained when both corrections are applied, confirming that the full FUSE procedure provides complementary benefits and stabilizes the flow integration process.

4.3.2 Importance of seasonal context

Similarly to DINCAE2 (Barth et al., 2022) and CRITER (Zupančič Muc et al., 2025), DIRECT incorporates seasonal context through the sine and cosine of the day-of-year d_t, which encode the annual cycle. To evaluate the contribution of this information, we trained a variant $\mathrm{DIRECT}_{\overline{\mathrm{DoY}}}$ that excludes DoY embeddings. As shown in Table 7, removing seasonal context increases $\mathrm{RMSE}_{\mathrm{mis}}$ from 0.234 to 0.242, corresponding to a relative degradation of about 3 %. Although the effect is modest, this indicates that DoY features provide useful information.

4.3.3 Influence of mask-aware conditioning

DIRECT incorporates one-hot encoded masks that distinguish between observed pixels, filled pixels originating from temporally shifted frames, and missing or land pixels. To evaluate the importance of these masks, we trained a variant $\mathrm{DIRECT}_{\overline{\mathrm{MASK}}}$ that excludes all mask information. As shown in Table 8, removing mask-aware conditioning increases $\mathrm{RMSE}_{\mathrm{mis}}$ from 0.234

Table 7. Performance of DIRECT and DIRECT$_{\overline{\text{DoY}}}$, which is without the day-of-year information. All of the reported reconstruction errors are in °C.

Variant	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
DIRECT$_{\overline{\text{DoY}}}$	0.121	0.242	0.000
DIRECT	**0.113**	**0.234**	**0.000**

Table 8. Performance of DIRECT and DIRECT$_{\overline{\text{MASK}}}$, which does not utilize one hot encoded masks. All of the reported reconstruction errors are in °C.

Variant	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
DIRECT$_{\overline{\text{MASK}}}$	0.120	0.240	0.001
DIRECT	**0.113**	**0.234**	**0.000**

to 0.240, a relative degradation of about 3 %. Although the effect is modest, the masks provide a consistent benefit by guiding the model to focus on reliable inputs.

4.3.4 One-hot encoding vs. offset masks

In the default DIRECT formulation, temporal origins of filled pixels are encoded using one-hot masks with three channels per auxiliary frame. As an alternative, we evaluate DIRECT$_{\text{OFFSETS}}$, which replaces the one-hot representation with a single binary mask and an integer offset map that encodes the temporal origin directly: -1, -2, and -3 indicate pixels sourced from $t-1$, $t-2$, and $t-3$, respectively, while a value of 0 marks missing or invalid pixels. The offset map for the future frame ($t+1$) is constructed in the same way, with values $+1$, $+2$, and $+3$. Table 9 confirms that both approaches perform similarly, with the one-hot representation giving a slight edge (less than 1 %). This suggests that compressed offset maps are a viable alternative when reducing input channels is desirable.

4.3.5 Influence of reconstruction ensemble size

Because DIRECT is a generative model, multiple reconstructions can be sampled for a single input by varying the initial noise seed. As per Section 2.2, these samples are averaged to produce the final reconstruction. Table 10 shows that averaging just

Table 9. Performance of DIRECT and DIRECT$_{\text{OFFSETS}}$. All of the reported reconstruction errors are in °C.

Variant	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
DIRECT$_{\text{OFFSETS}}$	0.115	0.236	0.000
DIRECT	**0.113**	**0.234**	**0.000**

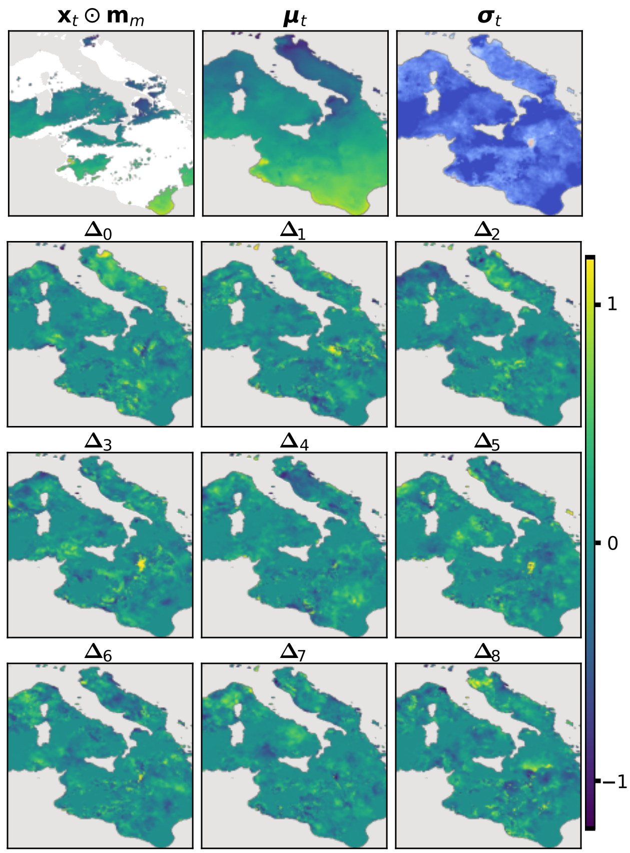

Figure 14. A visualization of a reconstruction ensemble. The first row shows the masked input, the ensemble mean ($\boldsymbol{\mu}_t$) and the per-pixel uncertainty ($\boldsymbol{\sigma}_t$). Rows 2–4 visualize the residual fields $\boldsymbol{\mu}_t - \hat{\mathbf{x}}_{t,n}$ for nine independent samples, illustrating the high-frequency structural diversity.

$N = 4$ reconstructions already reduces RMSE_{mis} by 18 % compared to $N = 1$, and crucially, this already outperforms the deterministic state-of-the-art CRITER. Our default $N = 16$ yields a 23 % improvement compared to $N = 1$. Although larger ensembles (e.g. $N = 32$) are comparable with our default, they also increase inference time, making 16 samples practical and effective.

Figure 14 illustrates the ensemble characteristics. The mean field provides a stable and accurate reconstruction, while the per-pixel standard deviation captures uncertainty in obscured regions. The residuals $\boldsymbol{\Delta}_n = \boldsymbol{\mu}_t - \hat{\mathbf{x}}_{t,n}$ (Rows 2–4) highlight the diverse high-frequency fluctuations present in individual samples.

4.3.6 Influence of multi-pass reconstruction

We evaluated a multi-pass inference strategy to test whether iteratively refining the temporal context improves the final reconstruction. In the default single-pass formulation (DIRECT_1), missing measurements in the temporal context (days $t-1$

Table 10. Performance impact of ensemble size N on reconstruction accuracy. All of the reported reconstruction errors are in °C.

Ensemble size N	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
$N = 1$	0.148	0.307	0.000
$N = 4$	0.125	0.251	0.000
$N = 8$	0.118	0.241	0.000
$N = 16$	**0.113**	**0.234**	**0.000**
$N = 32$	0.113	0.234	0.000
CRITER	0.127	0.255	0.017

Table 11. Influence of the multi-pass strategy (DIRECT_2), compared to the single-pass (DIRECT_1) and CRITER. All reported reconstruction errors are in °C. Med. denotes the Mediterranean dataset.

Dataset	Model	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
Med.	DIRECT_1	**0.113**	**0.234**	**0.000**
	DIRECT_2	0.113	0.235	0.001
	CRITER	0.127	0.255	0.017
Adriatic	DIRECT_1	**0.113**	**0.208**	**0.001**
	DIRECT_2	0.113	0.211	**0.001**
	CRITER	0.130	0.243	0.021
Atlantic	DIRECT_1	**0.363**	**0.489**	**0.001**
	DIRECT_2	0.375	0.498	**0.001**
	CRITER	0.391	0.518	0.036

and $t+1$) are opportunistically filled using the closest available observations. In the multi-pass formulation (DIRECT_{N_p}), we attempt to improve the central day's estimate by running the reconstruction of the entire time-series in multiple passes, substituting the missing values in the context days with the model's own predictions from the previous pass.

However, as summarized in Table 11, the two-pass strategy (DIRECT_2) yields performance comparable to, or slightly worse than, the single-pass baseline (DIRECT_1) across all datasets. In the Mediterranean and Adriatic regions, the reconstruction error in missing regions (RMSE_{mis}) increases slightly from 0.234 to 0.235 and from 0.208 to 0.211, respectively. On the Atlantic dataset—which is characterized by persistent, large-scale cloud cover—the error increases from 0.489 to 0.498.

These results indicate that substituting the initial fallback context with the model's own first-pass estimates propagates reconstruction uncertainties and errors rather than providing a cleaner conditioning signal.

Table 12. Influence of future temporal context. The past-only variant ($\text{DIRECT}_{\text{past}}$) is compared to the default model (DIRECT) and the CRITER baseline. All reported reconstruction errors are in °C. Med. denotes the Mediterranean dataset.

Dataset	Model	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
Med.	DIRECT	**0.113**	**0.234**	**0.000**
	$\text{DIRECT}_{\text{past}}$	0.117	0.241	0.001
	CRITER	0.127	0.255	0.017
Adriatic	DIRECT	**0.113**	**0.208**	**0.001**
	$\text{DIRECT}_{\text{past}}$	0.120	0.227	0.002
	CRITER	0.130	0.243	0.021
Atlantic	DIRECT	**0.363**	**0.489**	**0.001**
	$\text{DIRECT}_{\text{past}}$	0.369	**0.489**	0.002
	CRITER	0.391	0.518	0.036

4.3.7 Influence of future temporal context

In our default formulation, DIRECT utilizes a temporal context spanning both past and future observations ($t-1$ and $t+1$) to reconstruct the central day t. This bidirectional approach is standard for historical gap-filling tasks, as it allows the model to smoothly interpolate the evolution of ocean dynamics. However, for Near Real-Time (NRT) operational settings, future observations are strictly unavailable. To evaluate DIRECT's viability in such scenarios, we trained a past-only variant ($\text{DIRECT}_{\text{past}}$) that conditions the network solely on the preceding frame ($t-1$) and the central frame (t).

Table 12 compares $\text{DIRECT}_{\text{past}}$ against the default DIRECT. As expected, removing the future context leads to a slight degradation in performance compared to the default model. In the Mediterranean and Adriatic regions, the reconstruction error in missing regions (RMSE_{mis}) increases from 0.234 to 0.241, and from 0.208 to 0.227, respectively. On the Atlantic dataset, performance remains unchanged. This confirms that future information naturally aids the model in accurately resolving the temporal evolution of highly dynamic ocean events.

Crucially, despite lacking future temporal anchoring, $\text{DIRECT}_{\text{past}}$ still consistently outperforms the state-of-the-art CRITER across all three regions. This demonstrates that DIRECT's generative prior and observation-guided rectification remain highly effective even under strict constraints, making the architecture highly suitable for daily operational NRT applications.

4.3.8 Incorporating microwave observations

DIRECT operates on high-resolution infrared L3 observations, which are unavailable under cloud cover. Microwave radiometers, however, do retrieve SST through clouds, albeit at substantially coarser spatial resolution, and therefore offer a potential source of information beneath clouds. Because DIRECT ingests its spatio-temporal context through a lightweight input projection, additional observation sources can be added with minimal architectural change. To demonstrate this, we conducted a

Table 13. Effect of incorporating microwave (MW) observations on the Mediterranean dataset. All reported reconstruction errors are in °C. Med. denotes the Mediterranean dataset.

Dataset	Model	$\mathrm{RMSE_{all}}$	$\mathrm{RMSE_{mis}}$	$\mathrm{RMSE_{obs}}$
Med.	DIRECT	0.113	0.234	0.000
	$\mathrm{DIRECT_{MW}}$	0.112	0.230	0.001

preliminary experiment on the Mediterranean dataset using the daily microwave optimally-interpolated SST product MW OI SST v5.1 (Remote Sensing Systems, 2022), regridded and co-located to our domain. The microwave field is supplied as an additional input by embedding it through a separate convolution branch and summing it with the projected context, mirroring the treatment of $\mathbf{C}_{2t}$ (denoted $\mathrm{DIRECT_{MW}}$).

As shown in Table 13, the microwave variant leaves reconstruction accuracy essentially unchanged, with $\mathrm{RMSE_{mis}}$ marginally lower than the default model. We attribute the small effect to the coarse resolution of the microwave product relative to the infrared L3 fields. The purpose of the experiment is not to demonstrate an accuracy gain but to show that DIRECT is readily extensible to heterogeneous observation sources. A more careful treatment of the resolution mismatch and observation-specific uncertainty, together with an evaluation in regimes of persistent, large-scale cloud cover where cloud-penetrating observations are expected to be most valuable, is a promising direction that we leave to future work.

5 Discussion

DIRECT reconstructs an L3 satellite product, which is itself a retrieval carrying sensor noise and residual cloud contamination. Our metrics are computed against this product, thus the reported residuals combine reconstruction error with the observation error already present in the reference. The FUSE step re-injects the measured pixels at every integration step and thus inherits their noise, a deliberate trade-off in exchange for hard observation consistency. The model does not depend critically on this choice: the variant that disables observation re-injection ($\mathrm{DIRECT}_{\overline{\mathrm{OBS}}}$, Sec. 4.3.1) still reconstructs the missing regions to state-of-the-art accuracy, showing that performance is driven by the generative prior rather than the re-injection alone. Both training and evaluation use cloud masks from other days, which enables learning without cloud-free ground truth. Two consequences follow. Cloud occurrence is coupled to the atmospheric state and hence to SST variability, thus the missingness is not strictly independent of the field being reconstructed. And because we synthetically withhold observed pixels, the evaluation measures reconstruction skill against a known reference rather than validating under genuine cloud conditions. The temporal context uses a fixed three-day window, following prior evidence that most relevant temporal information lies within a few surrounding days. This is limited when informative observations fall outside the window, most notably during cloud systems that persist over many days. Cloud-penetrating microwave observations are well suited to these conditions, and our preliminary experiment (Sec. 4.3.8) shows that DIRECT incorporates such a source with minimal architectural change, making a more thorough integration targeted at persistent-cloud regimes a natural next step. Finally, two aspects bound reconstruction quality. The ensemble mean is stable and accurate, but averaging acts as a low-pass filter that attenuates part of the fine-scale variability present in individual samples. The uncertainty is calibrated with a single global variance offset, which removes the dominant under-dispersion and brings coverage close to nominal (Sec. 4.2.3) but cannot correct calibration that varies across space or across the tails. Spatially adaptive calibration, and a training-time scoring-rule objective in place of the post-hoc correction, are promising routes to sharper estimates.

6 Conclusion

We presented DIRECT, a diffusion-inspired generative model for reconstructing dense SST fields from partially observed satellite measurements. By formulating SST gap-filling as a conditional flow-matching task, DIRECT departs from deterministic reconstruction and instead produces an ensemble of physically plausible realizations, enabling both accurate reconstructions and spatially resolved uncertainty estimation. The model combines temporal context, seasonal conditioning, and observation-guided rectification within a single end-to-end framework, and is trained directly on sparse observations without cloud-free ground truth. Experimental results across Mediterranean, Adriatic, and Atlantic datasets show that DIRECT consistently outperforms current state-of-the-art methods, reducing reconstruction error in occluded regions by 6-14 % relative to the strongest existing method (CRITER (Zupančič Muc et al., 2025)) and by up to 62 % compared to other recent methods, while better preserving mesoscale spatial structure and producing well-calibrated uncertainty estimates. Its performance is robust across cloud-coverage regimes and holds under a past-only configuration, making it suitable for operational near-real-time use. Extending the architecture to multivariate oceanographic data is a promising direction toward a more holistic, physically constrained generative model.

Code and data availability. Implementation of DIRECT and the code to train and evaluate the model are available in the GitHub repository: https://github.com/G-Rovscek/DIRECT. We also include DIRECT weights pretrained on the *Mediterranean*, *Adriatic* and *Atlantic* datasets. The persistent version of our GitHub repository containing code under the MIT licence is available at https://doi.org/10.5281/zenodo.18875310 (Rovscek, 2026). All three used datasets can be found at https://doi.org/10.5281/zenodo.13923189 (Zupancic Muc et al., 2024).

Appendix A: Additional qualitative comparison figures

This section presents additional reconstruction figures obtained with DIRECT and CRITER (Zupančič Muc et al., 2025). A more detailed discussion can be found in Section 4.2.

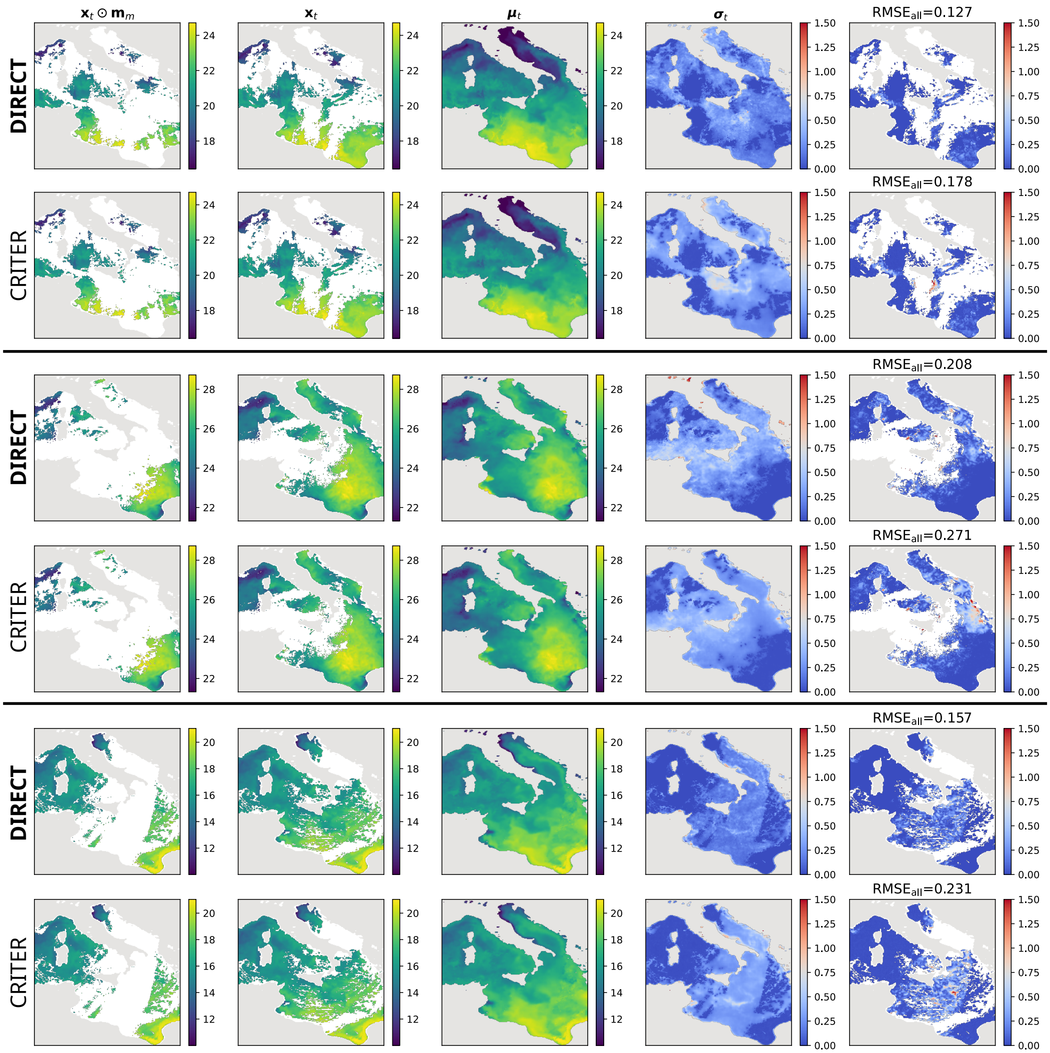

Figure A1. Comparison of DIRECT and CRITER reconstructions on the Mediterranean region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE$_{\text{all}}$). All values are in ˚C.

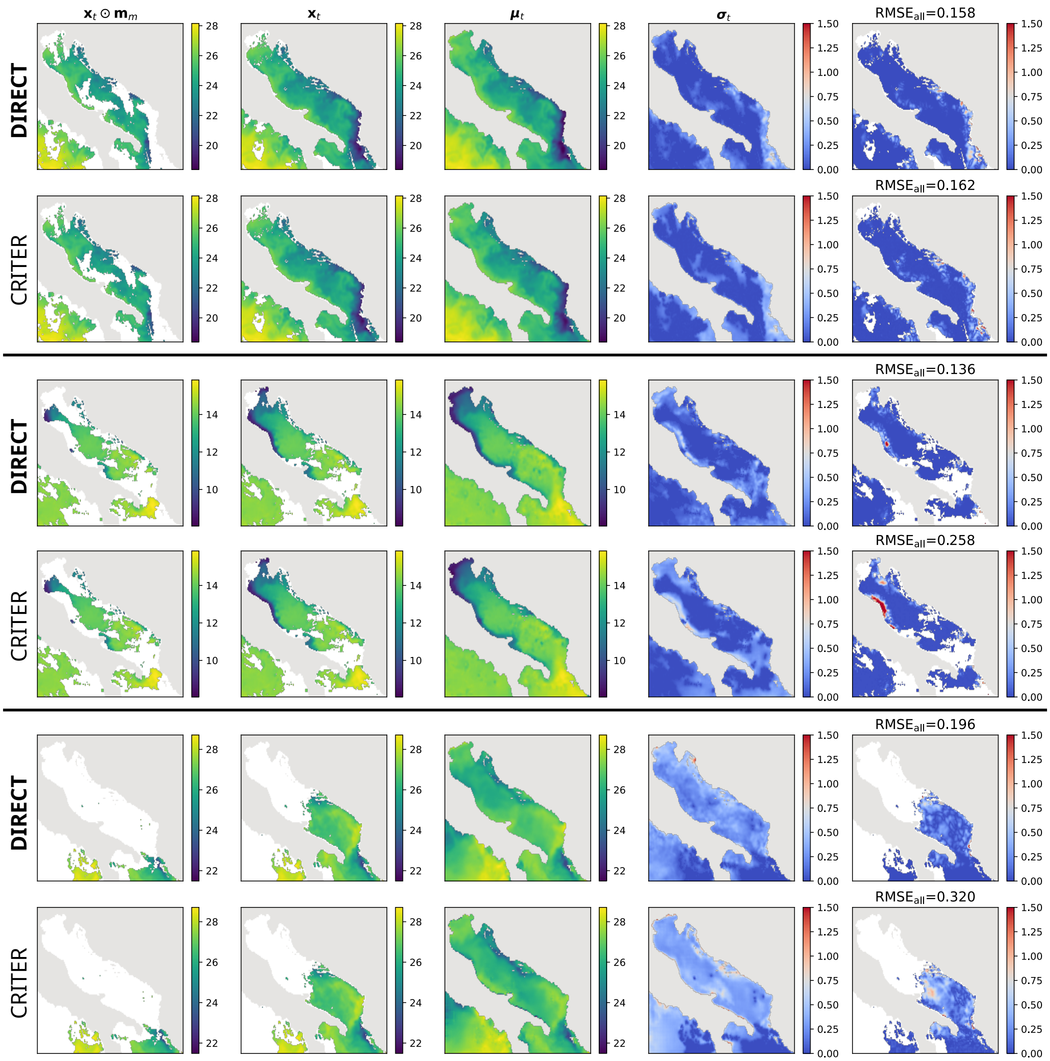

Figure A2. Comparison of DIRECT and CRITER reconstructions on the Adriatic region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE$_{\text{all}}$). All values are in °C.

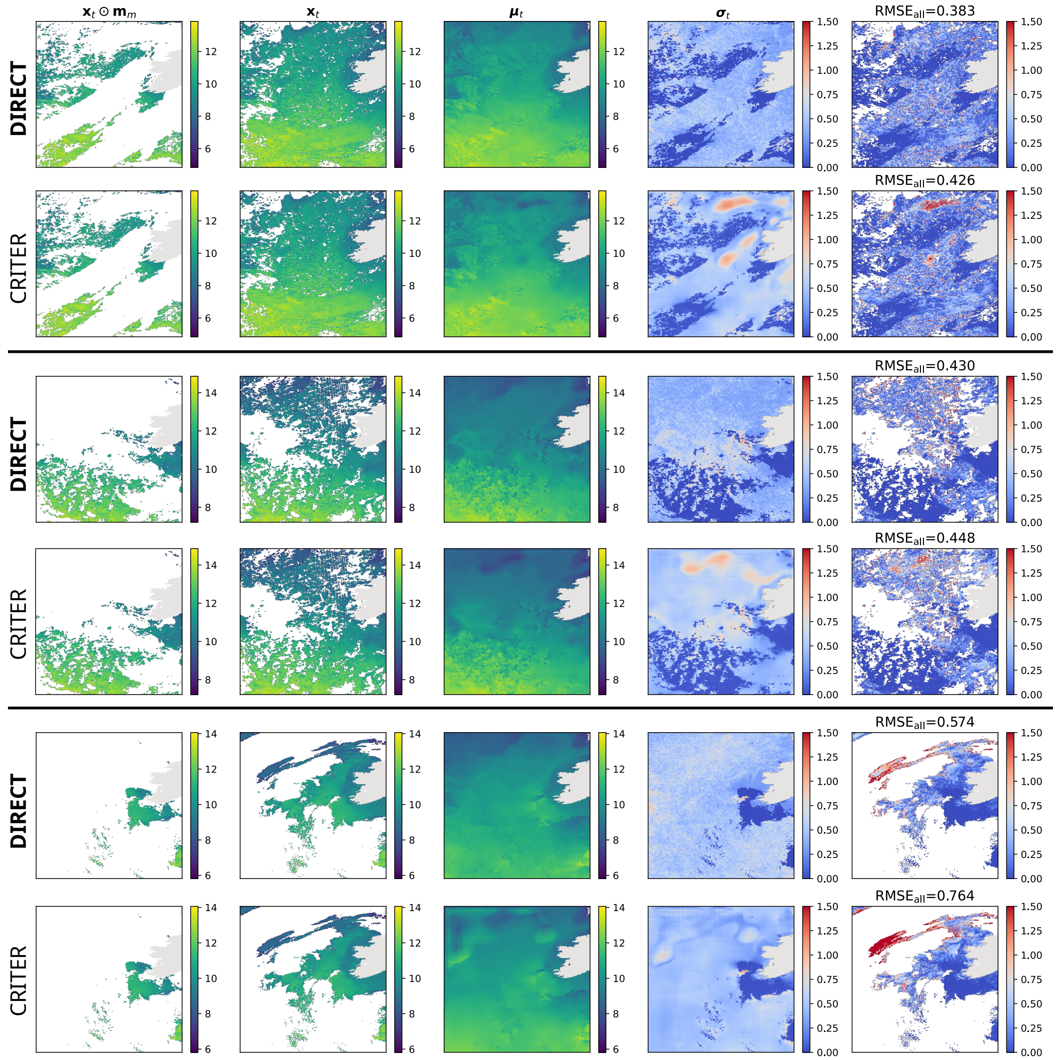

Figure A3. Comparison of DIRECT and CRITER reconstructions on the Atlantic region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE$_{\text{all}}$). All values are in ˚C.

Author contributions. GR developed the DIRECT code. MK led the machine learning part of the research. ML and AB contributed to the geophysical oceanographical part of the research. GR wrote the paper. All authors contributed to the final version of the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors would like to thank the Academic and Research Network of Slovenia - ARNES and the Slovenian National Supercomputing Network - SLING consortium (ARNES, EuroHPC Vega - IZUM) for making the research possible by using their supercomputer clusters. This study has been conducted using E.U. Copernicus Marine Service Information; https://doi.org/10.48670/moi-00171, https://doi.org/10.48670/moi-00310.

Financial support. This research was supported in part by ARIS program J2-2506 and projects P2-0214, and J2-60054. We acknowledge the support of the EC/EuroHPC JU and the Slovenian Ministry of HESI via the project SLAIF (grant number 101254461). Matjaž Ličer acknowledges the financial support from the Slovenian Research and Innovation Agency ARIS (contract no. P1-0237).

References

Alvera-Azcárate, A., Barth, A., Rixen, M., and Beckers, J.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature, Ocean Modelling, 9, 325–346, https://doi.org/https://doi.org/10.1016/j.ocemod.2004.08.001, 2005.

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, Geoscientific Model Development, 13, 1609–1622, 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, Geoscientific Model Development, 15, 2183–2196, https://doi.org/10.5194/gmd-15-2183-2022, 2022.

Barth, A., Brajard, J., Alvera-Azcárate, A., Mohamed, B., Troupin, C., and Beckers, J.-M.: Ensemble reconstruction of missing satellite data using a denoising diffusion model: application to chlorophyll a concentration in the Black Sea, Ocean Science, 20, 1567–1584, https://doi.org/10.5194/os-20-1567-2024, 2024.

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., and Zemp, M.: The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy, Bulletin of the American Meteorological Society, 95, 1431 – 1443, https://doi.org/10.1175/BAMS-D-13-00047.1, 2014.

Choo, M., Jung, S., Im, J., and Han, D.: CARE-SST: context-aware reconstruction diffusion model for sea surface temperature, ISPRS Journal of Photogrammetry and Remote Sensing, 220, 454–472, https://doi.org/https://doi.org/10.1016/j.isprsjprs.2025.01.001, 2025.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M.: Diffusion Models in Vision: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 10 850–10 869, https://doi.org/10.1109/TPAMI.2023.3261988, 2023.

Donlon, C., Robinson, I., Casey, K. S., Vazquez-Cuervo, J., Armstrong, E., Arino, O., Gentemann, C., May, D., LeBorgne, P., Piollé, J., Barton, I., Beggs, H., Poulter, D. J. S., Merchant, C. J., Bingham, A., Heinz, S., Harris, A., Wick, G., Emery, B., Minnett, P., Evans, R., Llewellyn-Jones, D., Mutlow, C., Reynolds, R. W., Kawamura, H., and Rayner, N.: The Global Ocean Data Assimilation Experiment High-resolution Sea Surface Temperature Pilot Project, Bulletin of the American Meteorological Society, 88, 1197 – 1214, https://doi.org/10.1175/BAMS-88-8-1197, 2007.

E.U. Copernicus Marine Service Information: Mediterranean Sea - High Resolution L3S Sea Surface Temperature Reprocessed, https://doi.org/10.48670/moi-00314, available from the Copernicus Marine Data Store (MDS), accessed 23-11-2024., 2024a.

E.U. Copernicus Marine Service Information: European North West Shelf/Iberia Biscay Irish Seas – High Resolution ODYSSEA Sea Surface Temperature Multi-sensor L3 Observations Reprocessed, https://doi.org/10.48670/moi-00311, available from the Copernicus Marine Data Store (MDS), accessed 23-11-2024., 2024b.

E.U. Copernicus Marine Service Information: Mediterranean Sea - High Resolution and Ultra High Resolution L3S Sea Surface Temperature, https://doi.org/10.48670/moi-00171, available from the Copernicus Marine Data Store (MDS), accessed 23-11-2024., 2024c.

E.U. Copernicus Marine Service Information (CMEMS): ESA SST CCI and C3S reprocessed sea surface temperature analyses, Marine Data Store (MDS), https://doi.org/10.48670/moi-00169, accessed 06-07-2026, 2026a.

E.U. Copernicus Marine Service Information (CMEMS): Global Ocean OSTIA Sea Surface Temperature and Sea Ice Reprocessed, Marine Data Store (MDS), https://doi.org/10.48670/moi-00168, accessed 06-07-2026, 2026b.

Garcia-Soto, C., Cheng, L., Caesar, L., Schmidtko, S., Jewett, E. B., Cheripka, A., Rigor, I., Caballero, A., Chiba, S., Báez, J. C., et al.: An overview of ocean climate change indicators: Sea surface temperature, ocean heat content, ocean pH, dissolved oxygen concentration, arctic sea ice extent, thickness and volume, sea level and strength of the AMOC (Atlantic Meridional Overturning Circulation), Frontiers in Marine Science, 8, 642 372, 2021.

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.

Harris, F.: On the use of windows for harmonic analysis with the discrete Fourier transform, Proceedings of the IEEE, 66, 51–83, https://doi.org/10.1109/PROC.1978.10837, 1978.

Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models, in: Advances in Neural Information Processing Systems, edited by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., vol. 33, pp. 6840–6851, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf, 2020.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, The Annals of Mathematical Statistics, 22, 79–86, http://www.jstor.org/stable/2236703, 1951.

Li, H., Wang, Y., Yang, K., Huang, G., Xia, X., Chen, Z., Tao, W., Lyu, C., Chen, L., Zhang, M., Hu, K., Gong, H., Fu, D., and Wang, L.: Probabilistic reconstruction of global sea surface temperature using generative diffusion models, https://arxiv.org/abs/2603.16272, 2026.

Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X., and Chen, Z.: Diffusion Models for Image Restoration and Enhancement – A Comprehensive Survey, https://arxiv.org/abs/2308.09388, 2023.

Li, Z., Wei, D., Zhang, X., Gao, Y., and Zhang, D.: A Daily High-Resolution Sea Surface Temperature Reconstruction Using an I-DINCAE and DNN Model Based on FY-3C Thermal Infrared Data, Remote Sensing, 16, https://doi.org/10.3390/rs16101745, 2024.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M.: Flow Matching for Generative Modeling, in: The Eleventh International Conference on Learning Representations, https://openreview.net/forum?id=PqvMRDCJT9t, 2023.

Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T. Q., Lopez-Paz, D., Ben-Hamu, H., and Gat, I.: Flow Matching Guide and Code, https://arxiv.org/abs/2412.06264, 2024.

Liu, H., Wang, Y., Qian, B., Wang, M., and Rui, Y.: Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8038–8047, 2024.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L.: RePaint: Inpainting Using Denoising Diffusion Probabilistic Models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11 461–11 471, 2022.

Matheron, G.: Principles of geostatistics, Economic geology, 58, 1246–1266, 1963.

Matheson, J. E. and Winkler, R. L.: Scoring Rules for Continuous Probability Distributions, Management Science, 22, 1087–1096, http://www.jstor.org/stable/2629907, 1976.

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, Volume 6 - 2019, https://doi.org/10.3389/fmars.2019.00420, 2019.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer, Proceedings of the AAAI Conference on Artificial Intelligence, 32, https://doi.org/10.1609/aaai.v32i1.11671, 2018.

Qian, Z., Liu, T., Bathiany, S., Yang, S., Hess, P., Bochow, N., Burmester, C., Gelbrecht, M., Groenke, B., and Boers, N.: Generative deep learning improves reconstruction of global historical climate records, https://arxiv.org/abs/2602.16515, 2026.

Rahman, R. and Rahaman, H.: Evaluation of sea surface temperature from ocean reanalysis products over the North Indian Ocean, Frontiers in Marine Science, Volume 11 - 2024, https://doi.org/10.3389/fmars.2024.1461696, 2024.

Remote Sensing Systems: Microwave OI SST Product, Version 5.1, https://www.remss.com/measurements/sea-surface-temperature/oisst-description/, data available at https://data.remss.com/SST/daily/mw/v05.1/netcdf/, accessed 13-07-2026, 2022.

Ricchi, A., Sangelantoni, L., Redaelli, G., Mazzarella, V., Montopoli, M., Miglietta, M. M., Tiesi, A., Mazzà, S., Rotunno, R., and Ferretti, R.: Impact of the SST and topography on the development of a large-hail storm event, on the Adriatic Sea, Atmospheric Research, 296, 107 078, 2023.

Rovscek, G.: Grega Rovscek/DIRECT: Release 1.0, https://doi.org/10.5281/zenodo.18875310, 2026.

Rovšček, G., Ličer, M., and Kristan, M.: Dense Spatiotemporal Reconstruction of Sea Surface Temperature with Conditional Flow Matching, in: Proceedings of the 29th Computer Vision Winter Workshop (CVWW), Jindřichův Hrec, Czech Republic, https://cmp.felk.cvut.cz/cvww2026/assets/pdfs/CVWW2026-43-final.pdf, 2026.

Senatore, A., Furnari, L., and Mendicino, G.: Impact of high-resolution sea surface temperature representation on the forecast of small Mediterranean catchments' hydrological responses to heavy precipitation, Hydrology and Earth System Sciences, 24, 269–291, https://doi.org/10.5194/hess-24-269-2020, 2020.

Song, Y., Lyu, P., Fei, B., Ling, F., Ouyang, W., and Bai, L.: ReconMOST: Multi-Layer Sea Temperature Reconstruction with Observations-Guided Diffusion, https://arxiv.org/abs/2506.10391, 2025.

Stoica, P., Moses, R. L., et al.: Spectral analysis of signals, vol. 452, Pearson Prentice Hall Upper Saddle River, NJ, 2005.

Taburet, G., Sanchez-Roman, A., Ballarotta, M., Pujol, M.-I., Legeais, J.-F., Fournier, F., Faugere, Y., and Dibarboure, G.: DUACS DT2018: 25 years of reprocessed sea level altimetry products, Ocean Science, 15, 1207–1224, https://doi.org/10.5194/os-15-1207-2019, 2019.

Wang, H., Han, J., Fan, W., Zhang, W., and Liu, H.: PhyDA: Physics-Guided Diffusion Models for Data Assimilation in Atmospheric Systems, https://arxiv.org/abs/2505.12882, 2025.

Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., and Jiang, Y.-G.: A survey on video diffusion models, ACM Computing Surveys, 57, 1–42, 2024.

Zupancic Muc, M., Zavrtanik, V., Barth, A., Alvera-Azcarate, A., Licer, M., and Kristan, M.: CRITER 1.0: Sea Surface Temperature Evaluation Datasets, https://doi.org/10.5281/zenodo.13923189, 2024.

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, Geoscientific Model Development, 18, 5549–5573, https://doi.org/10.5194/gmd-18-5549-2025, 2025.