Response to referee comment 1

(https://doi.org/10.5194/gmd-2024-208-RC1)

Review of "CRITER 1.0: A coarse reconstruction with iterative refinement network for sparse spatio-temporal satellite data" by Matjaž Zupancic Muc, Vitjan Zavrtanik, Alexander Barth, Aida Alvera-Azcarate, Matjaž Licer, and Matej Kristan

This manuscript is a description of a novel machine learning technique for gap filling of SST analysis in the presence of possibly significant missing (satellite) observational data. The reconstruction method uses high resolution, multi-sensor, binned where observations exist, L3S SST products from Copernicus Marine Service, and then uses a two stage approach machine learning technique to both fill the true missing data as well as further missing data removed from the L3S product to be used for training and validation. The analysis is then validated against this removed data, showing improvements over other methods -- primarily DINEOF of Alvera-Azcárate et al., (2005).

Firstly, I am not an expert on machine learning techniques, and therefore will offer limited comment of the techniques involved, but rather a potential user of improved SST analysis, and therefore offer comments more aligned with that perspective. This is one of my major points of commentary on the present manuscript: The paper as a whole is a rather technical description of the proposed method -- and rightly so. However, I believe some additional commentary on potential users of the system, and what benefits it might offer them should be addressed in the introduction. As it stands now, this is only addressed very briefly and casually in literally the first 4 lines of the introduction, after which the manuscript pivots to solely detailing the technical details. My additional major comment would be to better describe some of the terminology in the manuscript. The meaning of seemingly simple terminology, such as that used for variance, as well as deleted and visible regions, is likely inherently obvious to the authors, however, the interpretation of these terms by the reader could lead to some confusion. Some more detailed descriptions with regards to the used terminology may be necessary, even if this seems painfully obvious to the authors.

We thank the reviewer for their thoughtful and constructive feedback.Below, we provide a detailed, comment-by-comment response addressing each point raised.

Major comments

Comment 1: Not enough motivating background information in the introduction. Other than the first 4 lines of the introduction, no motivating information is provided as to why improved high-resolution SST reconstructions are necessary. While everyone would presumably like the best possible SST reconstruction, what applications would best benefit, and how might they benefit? Although more directed towards satellite capabilities than gap filling techniques, a review article such as "Observational Needs of Sea Surface Temperature" (https://doi.org/10.3389/fmars.2019.00420) would seem a good starting point for building motivation. Other articles exploring the use of improving the resolution of SST boundary conditions for numerical weather prediction could also prove useful. A quick search yielded me these two possibilities (10.1175/JCLI-3275.1, 10.5194/hess-24-269-2020). Presumably a more detailed background search would yield more.

Response 1: We thank the reviewer for pointing this out. We have now expanded on the motivation for our work and now the introductory paragraph reads:

1 Introduction

Infrared satellite sea surface temperature (SST) data are critical for ocean modeling, climate monitoring, fisheries management, and marine ecology (O'Carroll et al., 2019). On the one hand, the SST is a key boundary condition for atmospheric models extending from classical numerical weather prediction (Senatore et al., 2020; Chelton, 2005) to extreme storms (Ricchi et al., 2023) and climate variability (Garcia-Soto et al., 2021). In the ocean realm, continuous description of SST is vital for analyses of mesoscale (Bishop et al., 2017) and submesoscale baroclinic processes like fronts and eddies, but also for implementations of atmosphere-ocean couplings through turbulent heat fluxes (Strajnar et al., 2019; Ličer et al., 2016). Furthermore, vertical temperature profiles are a critical driver of heat, carbon and nutrient exchange between the surface and the deep ocean, and thus for a wide plethora of biogeochemical processes (Mogen et al., 2022) in the ocean surface boundary layer which depend on the temperatures above the pycnocline. Last but not least, SST is a key parameter for detection, mapping and analysis of

---

marine heatwaves (Hobday et al., 2016), and reconstructed satellite fields are imperative for determining the regional extent and intensity of such extreme events (Pastor and Khodayar, 2023; Darmaraki et al., 2019), which can have enormous impacts on acquaculture, fisheries and other aspects of economy (Gómez-Gras et al., 2021; Garrabou et al., 2022).

Such downstream applications therefore often require complete, dense measurement fields but cloud cover and sparse satellite coverage invariably lead to gappy and sparse data in both space and time. Reconstruction of gaps in the observations is therefore essential for a continuous description of ocean temperature fields and for many daily operational processes. These

We hope this addresses the reviewer's concerns.

Comment 2: Given that high resolution global NWP systems -- ECMWF's IFS is 9km (1/12º) -- better high resolution global SST products are also required. The SST reconstructions pursued in this manuscript are all regional (Mediterranean, Adriatic, North Atlantic). It is not mentioned whether it would be practical to scale the proposed technique to global domains, such as gap filling the Copernicus Marine Service 1/10º ODYSSEA L3 product.

Response 2: We appreciate the reviewer's critical point regarding the scalability of our method to global domains, such as the Copernicus Marine Service 1/10º ODYSSEA L3 product. While CRITER has demonstrated success in regional SST reconstruction, scaling to global resolutions is constrained by the memory demands of our model's global spatio-temporal attention mechanism. An obvious but not always available solution is to get access to a GPU cluster with enough memory to accommodate global domain training. The limitation could be circumvented, by classical techniques (similar to the ones found in typical implementations of optimal interpolation) such as tilling the domain (e.g., into 256 x 256 pixel regions) and processing each independently, followed by post-processing to mitigate boundary artifacts (i.e., applying overlapping tiles). Although this limits the exploitation of all available global context, it offers a practical pathway for scaling CRITER. Developing a memory-efficient, and possibly spatially iterative CRITER variant for global applications remains a challenging but promising direction for our future work.

Some terminology used in the manuscript, while seemingly obvious, on further contemplation the meaning and interpretation is not so obvious.

Comment 3: Uncertainty/Variance (σ^2): The uncertainty or variance outcome from the machine learning training process is introduced and summarized with the generic statement leading off section 3 in the opening 3 lines (ll. 82-84). This statement represents the only description of how this quantity, which plays a large role in the analysis of the techniques performance and skill over the rest of the manuscript. If possible a more detailed description of how this term is output or diagnosed from the machine learning process would be warranted. From a naive aspect, I would assume this variance, or uncertainty is the range of SST values that would lead to the same best fit outcome in the training process, but obviously, not enough information is given to confirm this. Furthermore, as detailed in the paper on "Observational Needs of Sea Surface Temperature" given above, and the outcome of many workshops on the needs required of SST observations and analysis, there is a strong need for estimates of uncertainty to accompany estimates of SST. The estimate of variance/uncertainty outcome from this technique seems well posed to fulfill this requirement -- if its definition is an adequate measure of this.

Response 3: We thank the reviewer for highlighting the need for a more detailed explanation of the uncertainty term "\sigma^2". We emphasize that the variance is not estimated as a fixed value during training. Rather, a network is *trained to predict* it from observations. In fact, we propose an iterative approach by the Iterative Refinement Module (IRM), whose two-channel output (reconstructed SST "\tilde{x}" and variance "\sigma^2"), is described in Section 3.2. At each pixel position "j" the IRM predicts a Gaussian distribution parametrized by the predicted

mean "\tilde{x}_(j)", and standard deviation "\sigma_(j)", following the approach of Barth et al. (2020). The model is trained to maximize the likelihood of the ground truth SST values “x_(j)” *hidden during training* (see loss function in Equation 4). This leads to the model assigning a higher variance “\sigma_(j)^2” in areas of higher expected reconstruction error. Importantly, “\sigma_(j)^2” thus represents the model’s predictive uncertainty. The variance prediction quality is validated in Section 5.3, where we show that “\sigma_(j)” correlates with empirical errors, which confirms its reliability as an uncertainty measure.

To clarify this, the following text (seen on the latexdiff below) has been added to Section 3.2,

3.2 Iterative refinement module (IRM)

To improve the reconstruction accuracy, the coarse reconstruction $\hat{\mathbf{x}}_t$ is refined by an iterative refinement module (IRM, Figure 2) through a sequence of residual improvements, producing the final reconstruction $\tilde{\mathbf{x}}$ and the corresponding uncertainty characterized by the variance $\boldsymbol{\sigma}^2$. Per pixel j, we model the reconstructed SST as a Gaussian distribution parameterized by predicted mean $\tilde{\mathbf{x}}_{(j)}$ and standard deviation $\boldsymbol{\sigma}_{(j)}$, following Barth et al. (2020). Note that $\boldsymbol{\sigma}^2$ emerges from training the model to minimize Equation 4, which penalizes over- and underestimation of the error variance $\boldsymbol{\sigma}^2$.

and the following to Section 3.3.

In the second stage, the parameters of CRM are fixed and only the parameters of IRM are trained. The training samples are generated as in CRM training, but since IRM produces the mean and variance of the reconstruction, the following negative log-likelihood loss is minimized as in DINCAE (Barth et al., 2020, 2022):

$$\mathcal{L}_{\mathrm{IRM}} = \frac{1}{|\mathbf{M}_t \odot \mathbf{M}_l|} \sum_{i=1}^{N} \left[\frac{(\mathbf{x}_{t(i)} - \tilde{\mathbf{x}}_{(i)})^2}{\boldsymbol{\sigma}^2_{(i)}} + \log(\boldsymbol{\sigma}^2_{(i)}) \right] \mathbf{M}_{t(i)} \mathbf{M}_{l(i)}, \tag{4}$$

where $\tilde{\mathbf{x}}$ and $\boldsymbol{\sigma}^2$ are the reconstruction and variance estimated after the last iteration in IRM, the summation goes over the N pixels in each of $\tilde{\mathbf{x}}$, $\boldsymbol{\sigma}^2$ and $\mathbf{x}_t$. This loss thus trains the model to assign higher variance to areas with greater expected reconstruction error. We validate the variance prediction quality in Section 4.4, by demonstrating its correlation with empirical errors.

Comment 4: The definition of variance becomes further confused with the introduction of scaled error (error divided by variance, l. 251, 3rd line of S5.3). While the authors again use symbol σ for the scaled variance, or more precisely, σ_ε, this is well identified. The confusion (for me) was then when scaled variance , $\sigma_\varepsilon << 1$ was compared with an idealized reconstruction where $\sigma_\varepsilon = 1$, this is casually referred to as an overestimate of the variance (ll. 261-262). It took me more than a few moments to eventually realize this was the scaled variance, with the actual variance being a divisor to this scaled variance -- and therefore scaled variance , $\sigma_\varepsilon < 1$, does indeed represent an overestimation of actual variance. At the risk of insulting some all knowing readers, but lifting up some of the slower to comprehend readers, please somehow remind the readers that this is the scaled variance which is divided by the actual variance -- and therefore the statement does actually make sense.

Response 4: We apologise for confusion and agree that the connection might not immediately be clear to even a skilled reader. To address the issue, we have added a sentence explaining how the value of standard deviation of the scaled error "\sigma_\epsilon" is interpreted before moving on to the analysis. Please see the corresponding latexdiff below.

4.4 Uncertainty estimation and bias analysis

CRITER and DINCAE2 estimate both the reconstruction of missing values and the associated uncertainty (i.e., the standard deviation) for each pixel. To assess the reliability of the estimated standard deviation, we employ the scaled error metric

~~$\epsilon_{(i)} = (\mathbf{x}_{(i)} - \tilde{\mathbf{x}}_{(i)})/\boldsymbol{\sigma}_{(i)},$~~

$$\epsilon_{(i)} = \frac{\mathbf{x}_{(i)} - \tilde{\mathbf{x}}_{(i)}}{\boldsymbol{\sigma}_{(i)}}, \tag{8}$$

as proposed by Barth et al. (2020). This metric quantifies the difference between the ground truth observation $\mathbf{x}_{(i)}$ and the reconstruction $\tilde{\mathbf{x}}_{(i)}$, normalized by the estimated standard deviation $\boldsymbol{\sigma}_{(i)}$, where i is the pixel index. We calculate the mean, μ_ϵ, and standard deviation, σ_ϵ, of the scaled error over the entire test set. Furthermore, we compute the bias, defined as the (non-normalized) mean difference between the ground truth observations and reconstructions. An ideal reconstruction method

would thus have the bias equal to zero (i.e., predicted values are not globally under or over estimated)~~and $\sigma_\epsilon = 1$.~~ and standard deviation of the scaled error σ_ϵ equal to one (i.e., per-pixel disparities match the predicted uncertainties). Standard deviation of the scaled error $\sigma_\epsilon < 1$ indicates that the predicted standard deviation $\boldsymbol{\sigma}$ is overestimated, while $\sigma_\epsilon > 1$ indicates that $\boldsymbol{\sigma}$ is underestimated.

Figure 10 displays the histogram of the scaled error metric $\epsilon_{(i)}$ for each test set, along with the corresponding Gaussian distribution, characterized by the estimated mean μ_ϵ and standard deviation σ_ϵ. The mean (μ_ϵ), standard deviation (σ_ϵ), and the bias for each dataset are provided in Table 2. Notably, CRITER moderately underestimates the standard deviation, with standard deviation of the scaled error σ_ϵ values of $1.116, 1.082$, and 1.156 on the Mediterranean, Adriatic, and Atlantic datasets, respectively, ranging from 8% to 16%. In contrast, on average, DINCAE2 significantly overestimates the standard deviation, with σ_ϵ values of $0.334, 0.996$, and 0.801 across the three datasets. The over-estimation thus ranges from as little as 0.4% to substantial over-estimates of 66%. CRITER consistently exhibits a very low bias (~~in order of 10^{-2}~~ of the order of 0.01 °C or lower) over all datasets. Furthermore, CRITER exhibits a significantly smaller bias on the Mediterranean and Adriatic datasets than DINCAE2, whereas DINCAE2 achieves a smaller bias on the Atlantic dataset. Note that, on the Adriatic dataset, DINCAE2 exhibits $18\times$ larger bias than CRITER.

Comment 5: Visible and Deleted regions: The definition of deleted regions seem relatively obvious: The regions where SST observations have artificially been removed from the L3 product. However, the definition of visible, sometimes referred to as observed, regions seems less definite: Is it the fully observed region in the L3 SST before deletion, or the observed region in the L3 SST after removal of the deleted regions? Please provide a precise definition of deleted and visible regions.

Response 5: Thank you for pointing this out. Deleted regions correspond to observations in the L3 SST product that were *artificially removed* by simulated clouds and thus withheld during the

training. Visible regions refer to the remaining observations in the L3 product after the removal of these deleted regions. To make this clearer, we have added the definition of these regions to Section 4.2 as shown on the latexdiff below.

For additional insights we ~~also~~ compute the RMSE separately ~~on deleted and on the visible regions~~ for (i) deleted regions, corresponding to observations artificially removed by simulated clouds in the L3 SST product and thus withheld during the training, and (ii) visible regions, corresponding to remaining observations post-deletion in $\mathbf{x}_t$ ~~as follows.~~.

The reconstruction error of deleted regions is defined as

$$\mathrm{RMSE}_{\mathrm{mis}} = \sqrt{\frac{\sum_{i=1}^{N}\left[(\mathbf{x}_{t(i)} - \tilde{\mathbf{x}}_{(i)})^2 \mathbf{M}_{t(i)} \mathbf{M}_{l(i)} (\mathbf{1} - \mathbf{M}_{m(i)})\right]}{|\mathbf{M}_t \odot \mathbf{M}_l \odot (\mathbf{1} - \mathbf{M}_m)|}}, \tag{6}$$

$\mathbf{M}_m$ is the mask of deleted regions, and $|\mathbf{M}_t \odot \mathbf{M}_l \odot (\mathbf{1} - \mathbf{M}_m)|$ denotes the number of deleted ground truth measurements.

The reconstruction error of visible regions is defined as

$$\mathrm{RMSE}_{\mathrm{vis}} = \sqrt{\frac{\sum_{i=1}^{N}\left[(\mathbf{x}_{t(i)} - \tilde{\mathbf{x}}_{(i)})^2 \mathbf{M}_{t(i)} \mathbf{M}_{l(i)} \mathbf{M}_{m(i)}\right]}{|\mathbf{M}_t \odot \mathbf{M}_l \odot \mathbf{M}_m|}}, \tag{7}$$

Typographic and style comments:

Comment 6: Section 4 Results (l. 168) is empty?

Response 6: Thank you for pointing out this oversight. The Results section was supposed to be followed by an Implementation details **subsection**. However, we've mistakenly labeled it as a **section**, which is why the results appeared empty. We've fixed this mistake as shown on the latexdiff below.

4 Results

5 ~~Implementation details~~

4.1 Implementation details

CRITER is implemented using the PyTorch library (Paszke et al., 2017) and trained on an NVIDIA Tesla V100 GPU. CRM block is trained with batch size of 8 using the AdamW optimizer with a learning rate $\alpha = 3e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.95$ for 60 epochs (warm-up period), then with a cosine decay scheduler (Loshchilov and Hutter, 2016) with step size 30 for another

---

140 epochs. In the next phase IRM block is trained using the pre-trained CRM with fixed parameters. We train IRM using the Adam optimizer, with $\alpha = 3e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ for 300 epochs, using a step learning rate scheduler with step size 50 and multiplicative factor $\gamma = 0.5$.

Comment 7: Figures 4-6, B1-B3: Limits on σ and rmse. The colour scale limits on σ and rmse seem to be all automatically generated. This is a hindrance to both comparing between techniques (CRITER/DINCAE2) and comparing over-dispersive and under-dispersive regions (σ vs rmse). Although I realize this will often lead to regions of colour saturation, I would strive (at

least for individual scenarios) to have the colour scale range identical between CRITER/DINCAE2 results and between σ and rmse (preferably with the zero value always represented). This would likely enhance your ability to discuss the results in the text, and by setting the scales for σ and rmse identically, it would then allow you to connect the results in Section 5.3 with the earlier results -- for instance, you would easily be able to identify regions where DINCAE2 has insufficient variance compared to RMSE, and vice versa for CRITER).

Response 7:

We thank the reviewer for the helpful suggestion. We have updated the figures accordingly: a common color scale is now used for both σ and RMSE, and across CRITER and DINCAE2 results. However, we encountered challenges due to the significantly different distributions of the two methods. To improve readability, we limited the color scale to the 0th–90th percentile of the data and selected a new colormap. We have also reduced the number of samples displayed and increased the size of each image to better highlight differences in the SST fields.

These adjustments provide a clearer visual comparison for the Adriatic and Atlantic datasets. However, for the Mediterranean dataset, DINCAE2's σ values are confined to a narrow range, and as a result, the image appears nearly uniform due to color scale saturation**.** Unfortunately, we were unable to resolve this without compromising comparability across the other scenarios.

Please see the updated figures below.

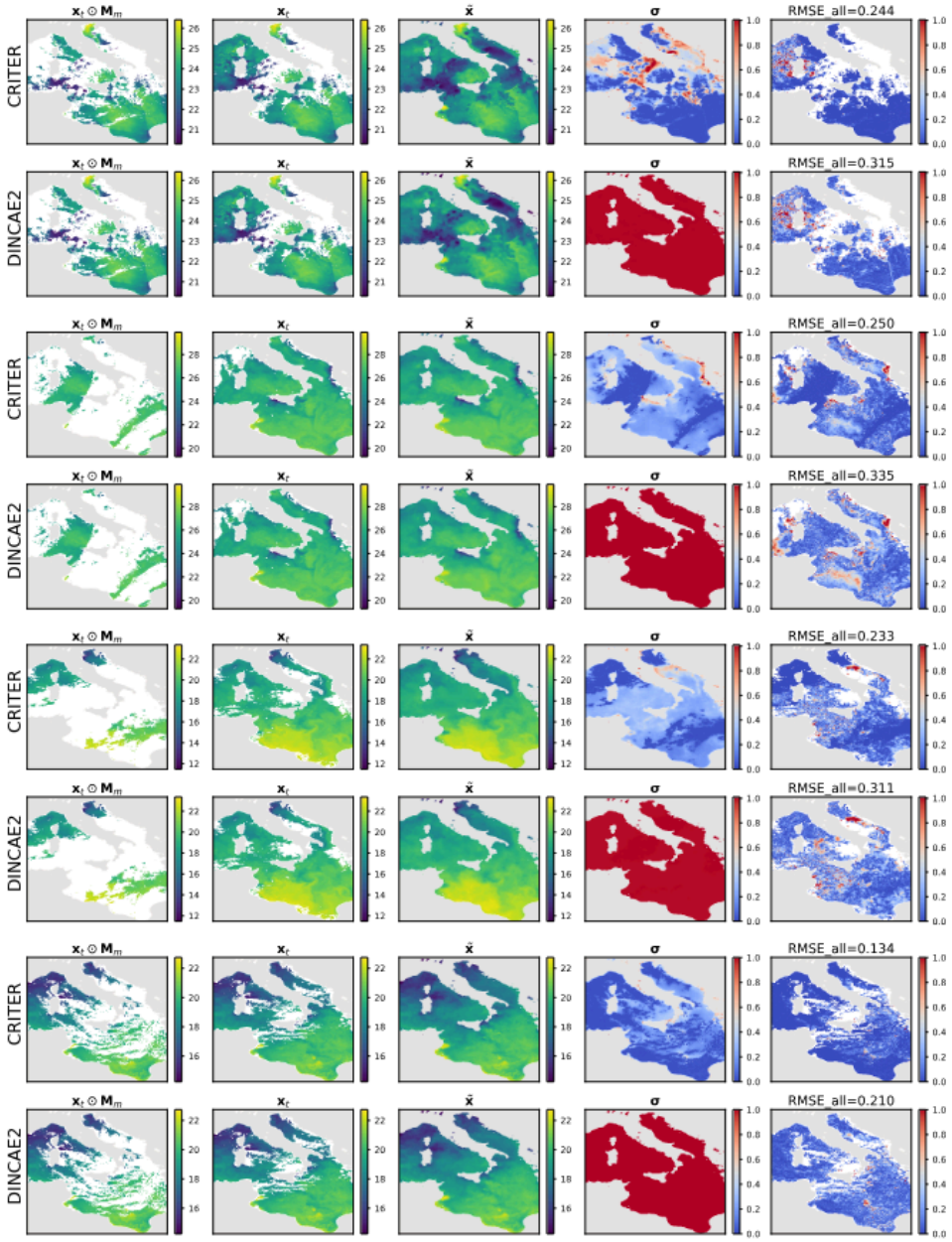

Figure 4. Comparison of sea surface temperature (SST) reconstructions generated by CRITER and DINCAE2 on the Mediterranean dataset. The columns display: (1) the original SST field with simulated missing values, (2) the original SST field, (3, 4) full reconstruction of the SST field and the associated standard deviation, and (5) the absolute error map, highlighting the differences between the original and reconstructed fields. All panel values are in °C. Note that color scales for $\boldsymbol{\sigma}$ and $\mathrm{RMSE}_{\mathrm{all}}$ are truncated at the 90th percentile of the data to improve visibility.

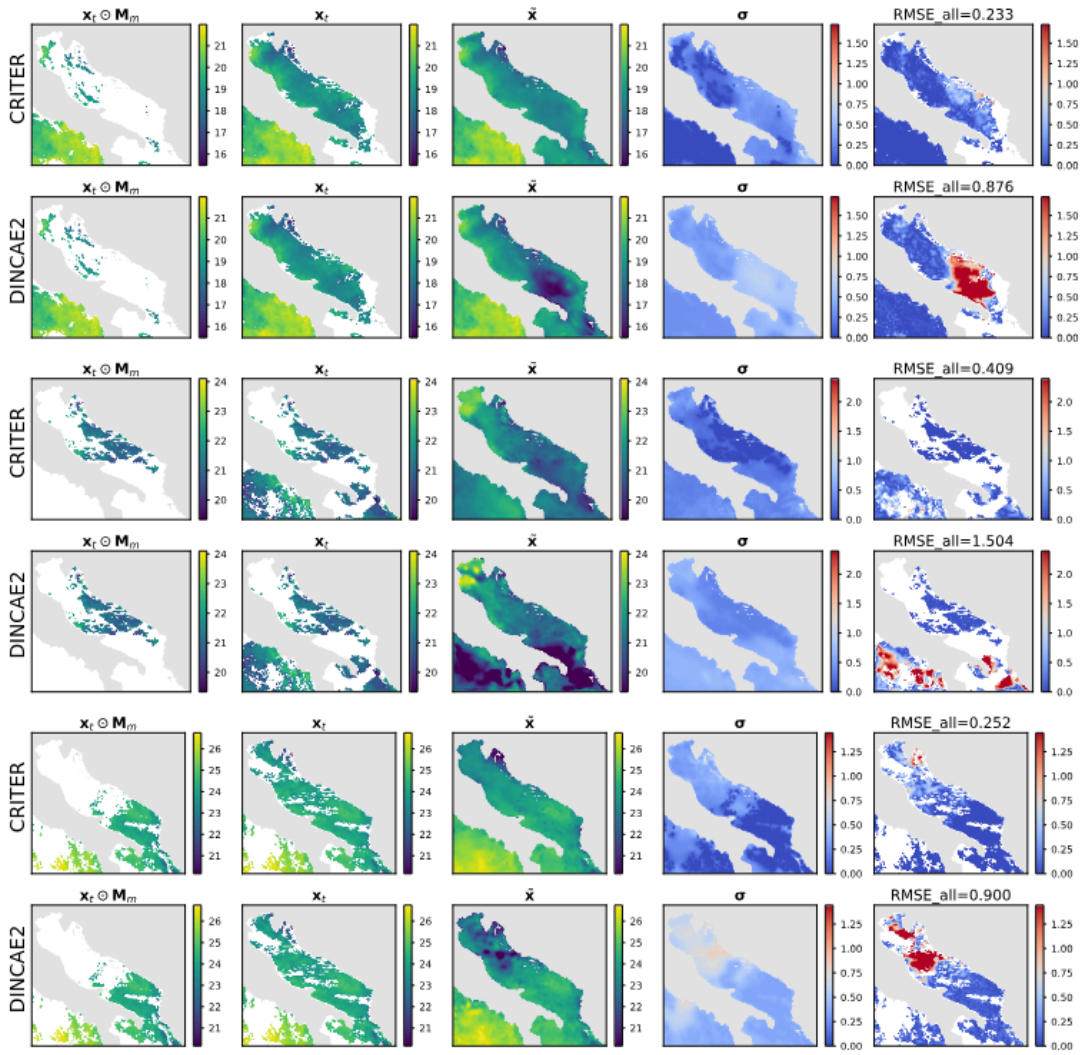

Figure 5. Same as Figure 4, but for the Adriatic domain.

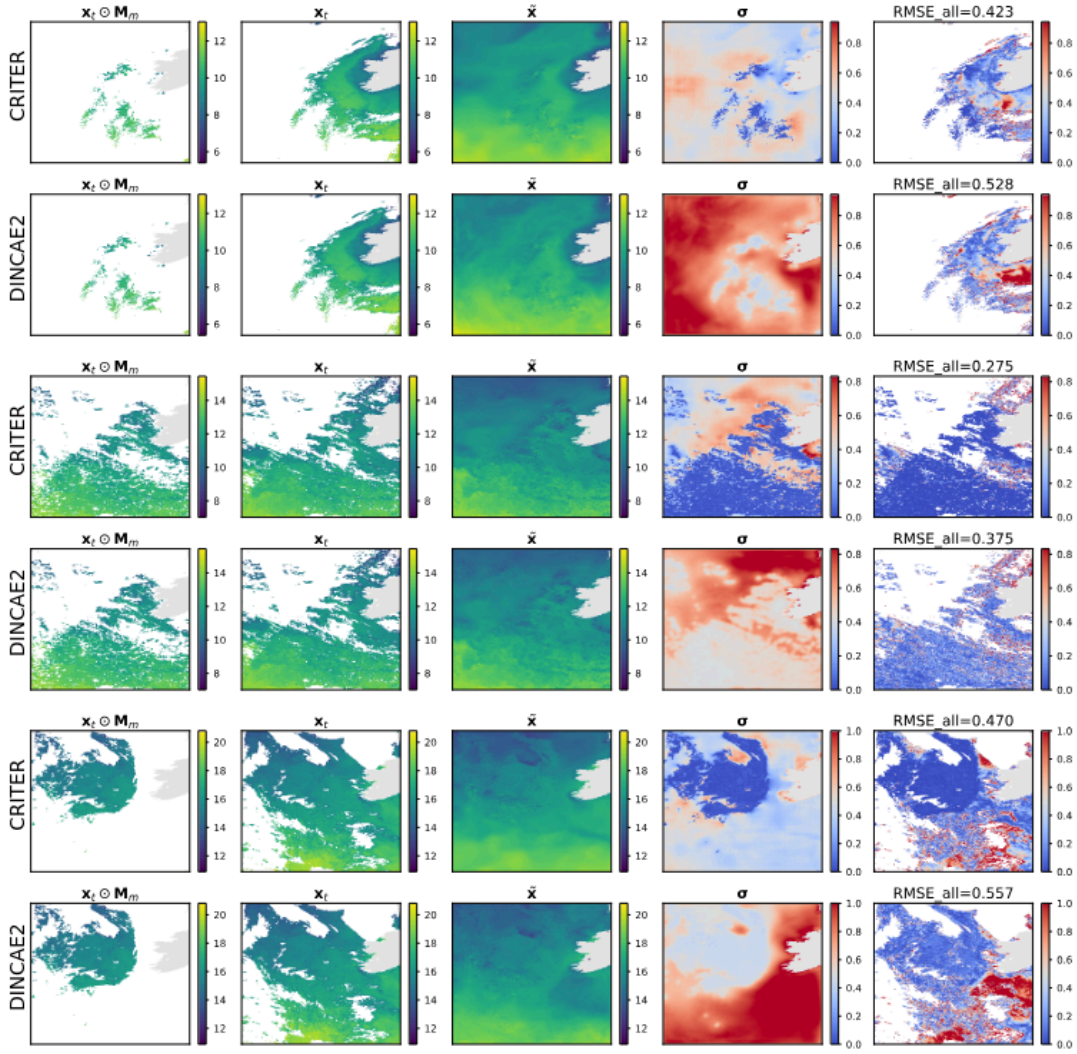

Figure 6. Same as Figure 4, but for the Atlantic domain.

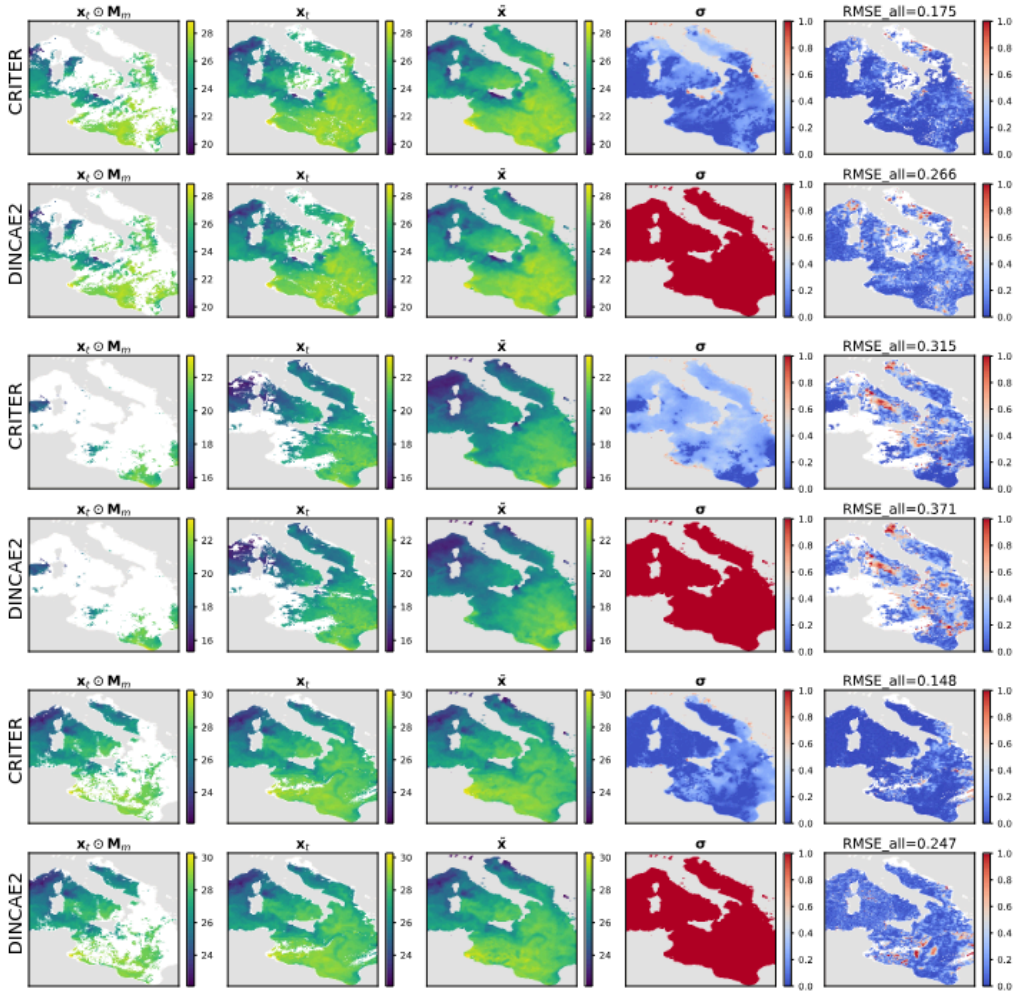

Figure B1. Same as Figure 4, on different samples.

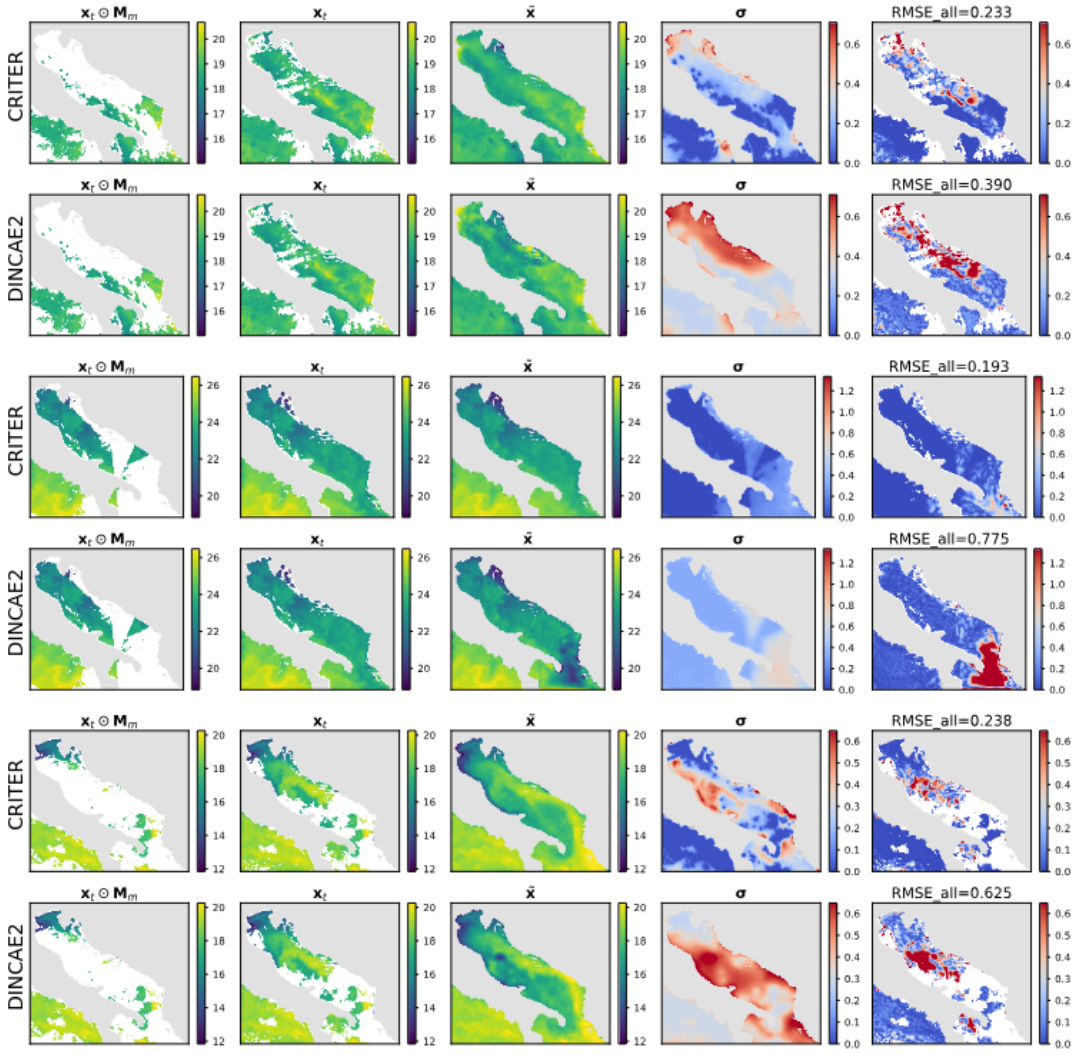

Figure B2. Same as Figure 4, but for the Adriatic domain.

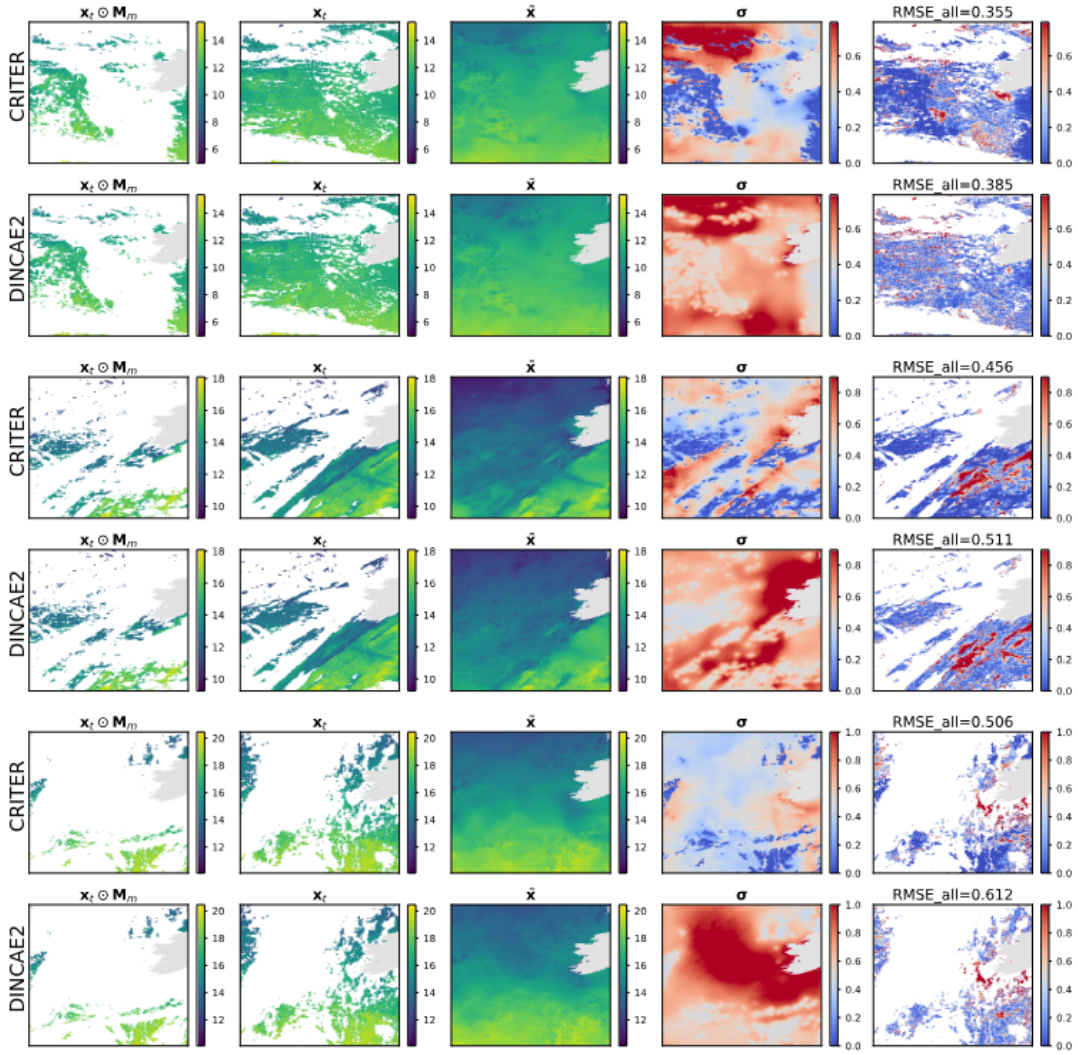

Figure B3. Same as Figure 4, but for the Atlantic domain.

Comment 8: l. 263 : in order of $10^{-2\circ}$ or lower. Not sure what is meant by to the power of -2o? Possible typographical error. "In order of" is more conventionally referred to as "Of the order of"

Response 8: Thank you for catching this. We've added the missing "C" symbol for Celsius in line 263 and corrected the phrasing to the conventional "of the order of".

with σ_ϵ values of $0.334, 0.996$, and 0.801 across the three datasets. The over-estimation thus ranges from as little as 0.4%
335 to substantial over-estimates of 66%. CRITER consistently exhibits a very low bias (~~in order of $10^{-2\circ}$~~ of the order of 0.01 °C or lower) over all datasets. Furthermore, CRITER exhibits a significantly smaller bias on the Mediterranean and Adriatic datasets than DINCAE2, whereas DINCAE2 achieves a smaller bias on the Atlantic dataset. Note that, on the Adriatic dataset, DINCAE2 exhibits $18\times$ larger bias than CRITER.

Further Comments:

Comments outside scope of manuscript that may be worthy of at least some discussion.

Comment 9: As already mentioned, I do not know much about technical details of machine learning techniques. But what I do know is that the techniques are relatively agnostic to the physical relationship between the inputs and outputs. In this study, one inputs binned temperature observations and outputs the full field temperatures. The input L3S products used in the study have undergone a variety of processing (radiance algorithms, bias corrections) to produce a binned multi-sensor temperature. Would this technique be generalizable to the underlying radiances, complicated by requiring different training for each instrument? The advantage might, however, be better instrument bias corrections and a further reduction in error?

Response 9: We appreciate the reviewer's insight into the potential benefits of applying CRITERIA to raw radiances. CRITER is designed for gridded SST data (L3/L3S products) structured as matrices representing spatially binned measurements. Raw radiances, however, are non-gridded. Adapting CRITER would require significant architectural modifications; for instance, replacing convolutional and transformer-based modules with methods suited for irregularly sampled data (e.g., point cloud networks).

However, even if gridded radiances product would have been available, the training objective could be twofold:

(i) radiance reconstruction: if the model is trained on raw radiances, its output would typically be reconstructed radiance, not SST

(ii) temperature estimation: to instead produce SST estimates, the training data must include paired radiance-temperature samples. However, in the latter case, if temperature is derived from radiance via fixed equations, any errors in this transformation would be learned by the model, potentially propagating and amplifying biases. While using direct temperature measurements (e.g., in situ data) as targets could resolve this, obtaining sufficient amounts of independent in situ data would be very challenging. Therefore, although adapting CRITER to raw radiances might leverage fine-grained information and reduce reconstruction errors, these architectural and data challenges currently render the approach infeasible.

All these trajectories are relevant for future work and we have amended the Conclusions of the paper to include these research possibilities.

Comment 10: As mentioned in Major Comment #1: Is this scalable to a global analysis?

Response 10: As mentioned in **Response 2**: We appreciate the reviewer's critical point regarding the scalability of our method to global domains, such as the Copernicus Marine Service 1/10° ODYSSEA L3 product. While CRITER has demonstrated success in regional SST reconstruction, scaling to global resolutions is constrained by the memory demands of our model's global spatio-temporal attention mechanism. An obvious but not always available solution is to get access to a GPU cluster with enough memory to accommodate global domain training. The limitation could be circumvented, by classical techniques (similar to the ones found in typical implementations of optimal interpolation) such as tilling the domain (e.g., into 256 x 256 pixel regions) and processing each independently, followed by post-processing to mitigate boundary artifacts (i.e., applying overlapping tiles). Although this limits the exploitation of all available global context, it offers a practical pathway for scaling CRITER. Developing a memory-efficient, and possibly spatially iterative CRITER variant for global applications remains a challenging but promising direction for our future work.

Comment 11: It could be interesting to apply a (spatial) spectral analysis on the results and underlying inputs, which admittedly would likely require large cloud free areas, at least for analyzing the spectral characteristics of the inputs. Do the wavelength characteristics of the CRITER and DINCAE2 results differ, and how do they compare to the original wavelength characteristics of the binned SST L3S products: Are certain wavelengths removed and/or enhanced?

Response 11: We thank the reviewer for the suggestion and have added Section 4.3.2 ("Spatial Spectral Analysis"), and Appendix C1 ("Extended Spatial Spectral Analysis") comparing the Power Spectral Density (PSD) of ground-truth observations against reconstructions from CRITER and DINCAE2. Please see the latex below.

4.3.2 Spatial Spectral Analysis

We conduct spatial spectral analysis by comparing the Power Spectral Density (PSD) of ground-truth observations against reconstructions from CRITER and DINCAE2, focusing on the Ionian Sea region due to its significant SST variability.

First, we identify observation fields with maximum number of known measurements within the ROI (Region Of Interest) and compute their PSDs over the ROI. Following Fanelli et al. (2024), we compute PSD using FFT with a Blackman-Harris window. We then sample 30 cloud masks with distinct coverage over the ROI, with the fraction of missing values ranging from 50% to 98%. For each mask, we simulate missing data in the observation fields, reconstruct them using both methods, and compute PSD over the reconstructed ROI.

Figure 7 shows an observation sequence with few available easurements. Both methods maintain PSD values near the target at low wavenumbers, indicating comparable low-frequency reconstruction. For wavenumbers $k \geq 4 \frac{\text{cycles}}{\text{deg}}$, however, CRITER's PSD remains closer to the target than DINCAE2's, demonstrating its superior ability to resolve high-frequency components. Figure 8 depicts a case with more measurements, where both methods generally align closer to the target. Nevertheless, CRITER still outperforms DINCAE2 at high wavenumbers ($k \geq 5 \frac{\text{cycles}}{\text{deg}}$). Additional results are provided in Appendix C1.

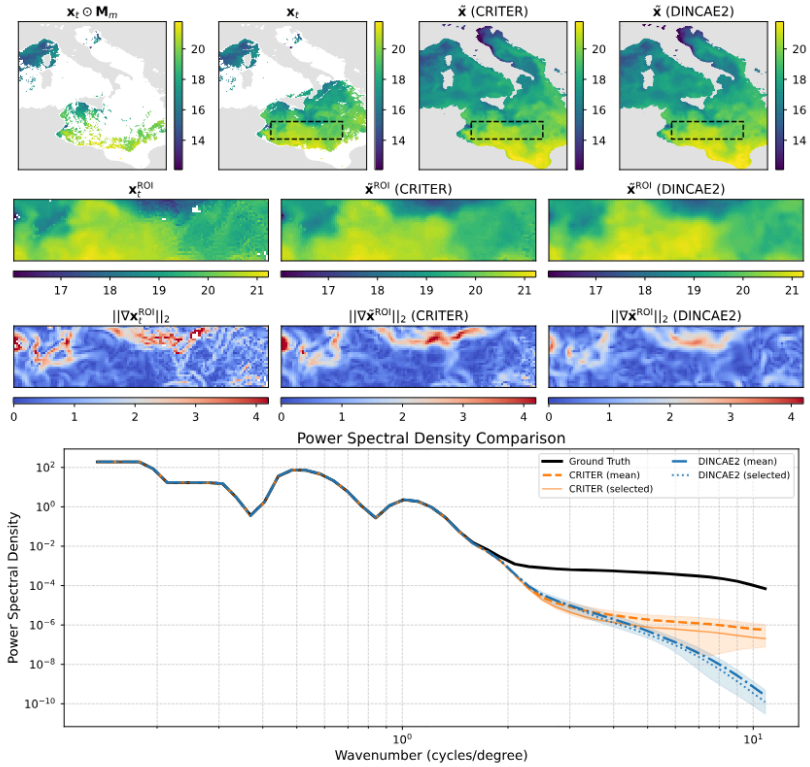

Figure 7. Visualization of reconstruction performance: Row 1 shows the full fields (left-to-right: Masked SST, Target SST, CRITER reconstruction, DINCAE2 reconstruction) with the Region of Interest (ROI) marked by a black-dashed rectangle. Row 2 displays the corresponding ROI fields: Target SST, CRITER reconstruction, and DINCAE2 reconstruction. Row 3 presents gradient magnitudes within the ROI for target, CRITER, and DINCAE2 outputs. Row 4 compares Power Spectral Densities: Target ROI (black), CRITER mean $\pm$ std (orange band), DINCAE2 mean $\pm$ std (blue band), with solid orange and dotted blue lines showing CRITER's and DINCAE2's PSDs for the selected example.

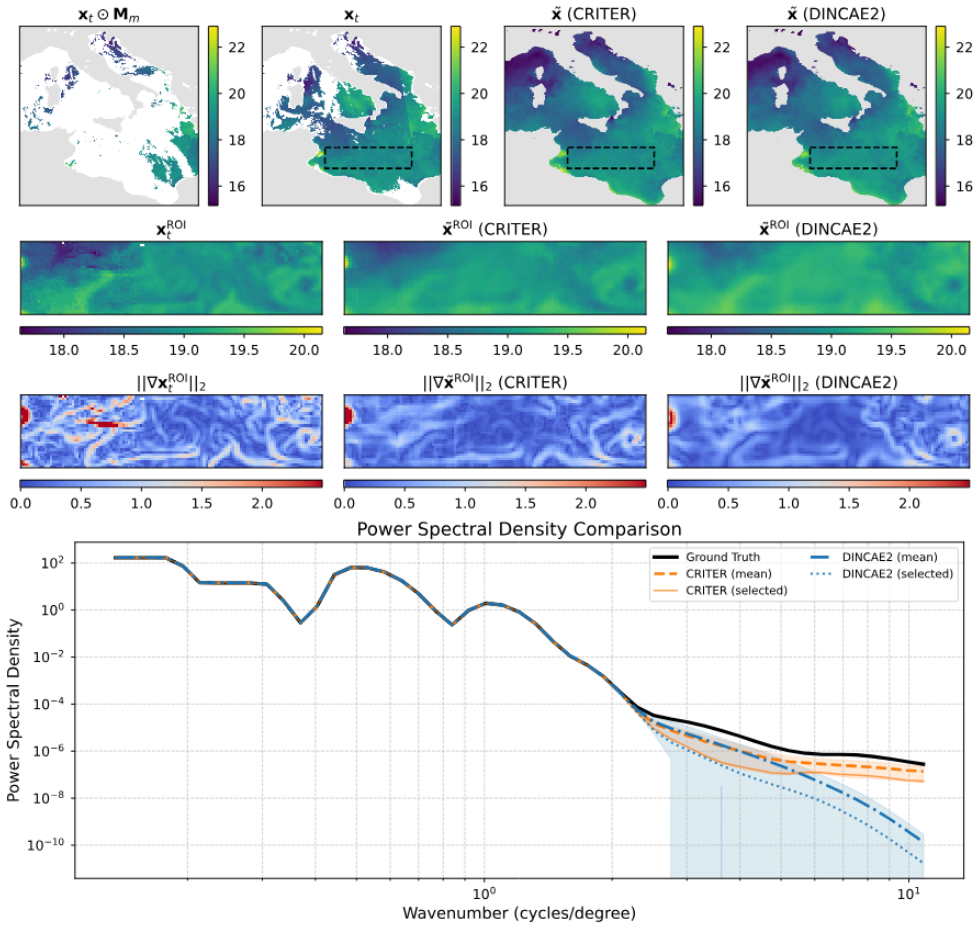

Figure 8. Same as Figure 7, but for another sample.

Appendix C

C1 Extended Spatial Spectral Analysis

This section presents supplementary power spectral density (PSD) comparisons. Figure C1 shows a challenging case with sparse measurements where CRITER's PSD remains closer to the target (on average) for wavenumbers $k \geq 4 \frac{\text{cycles}}{\text{deg}}$. Figure C2 depicts a high-measurement scenario featuring a failure case for CRITER: minor noise amplification beyond $k \geq 5 \frac{\text{cycles}}{\text{deg}}$. A similar issue occurs with DINCAE2, but in a different wavenumber band: Figure C3 shows significant noise amplification within $k \in [2, 4] \frac{\text{cycles}}{\text{deg}}$. For a detailed discussion of the comparison, refer to Section 4.3.2.

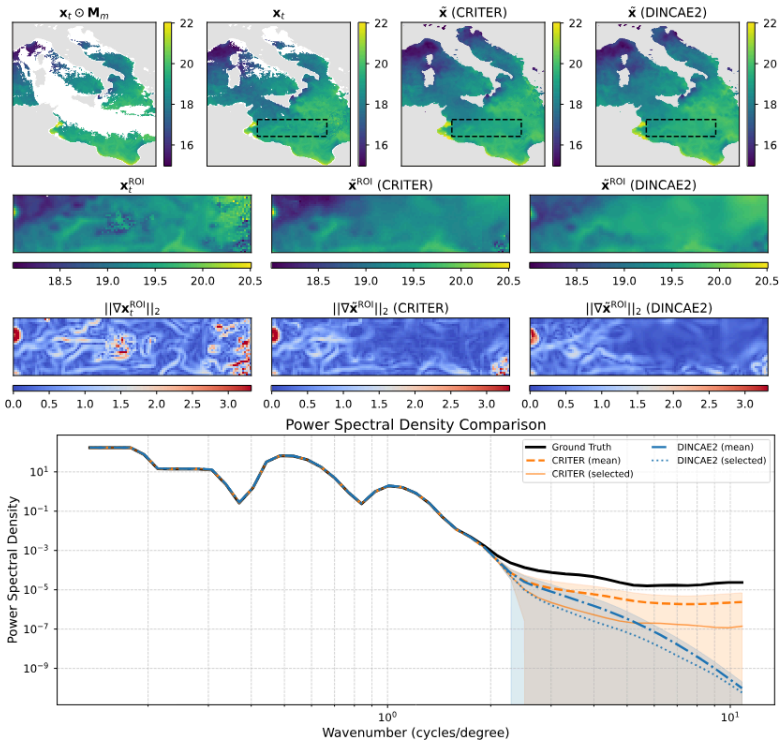

Figure C1. Same as Figure 7, but for a different sample.

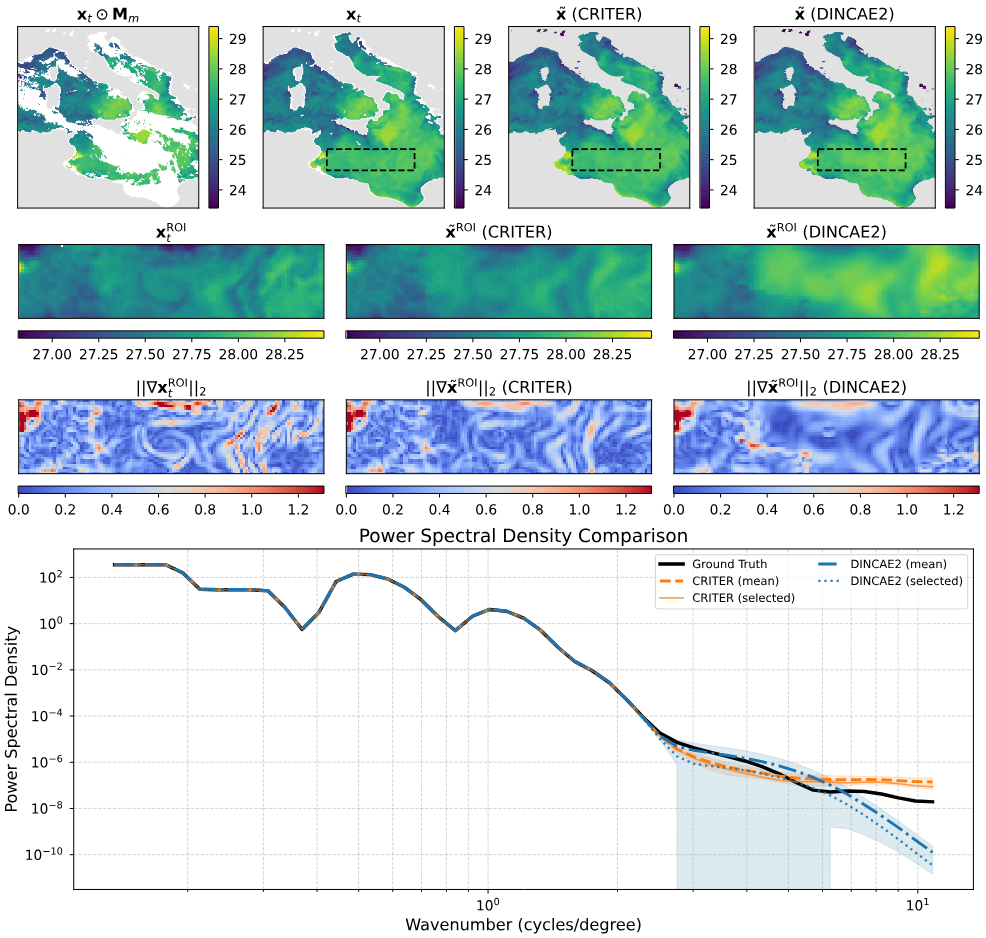

Figure C2. Same as Figure 7, but for a different sample.

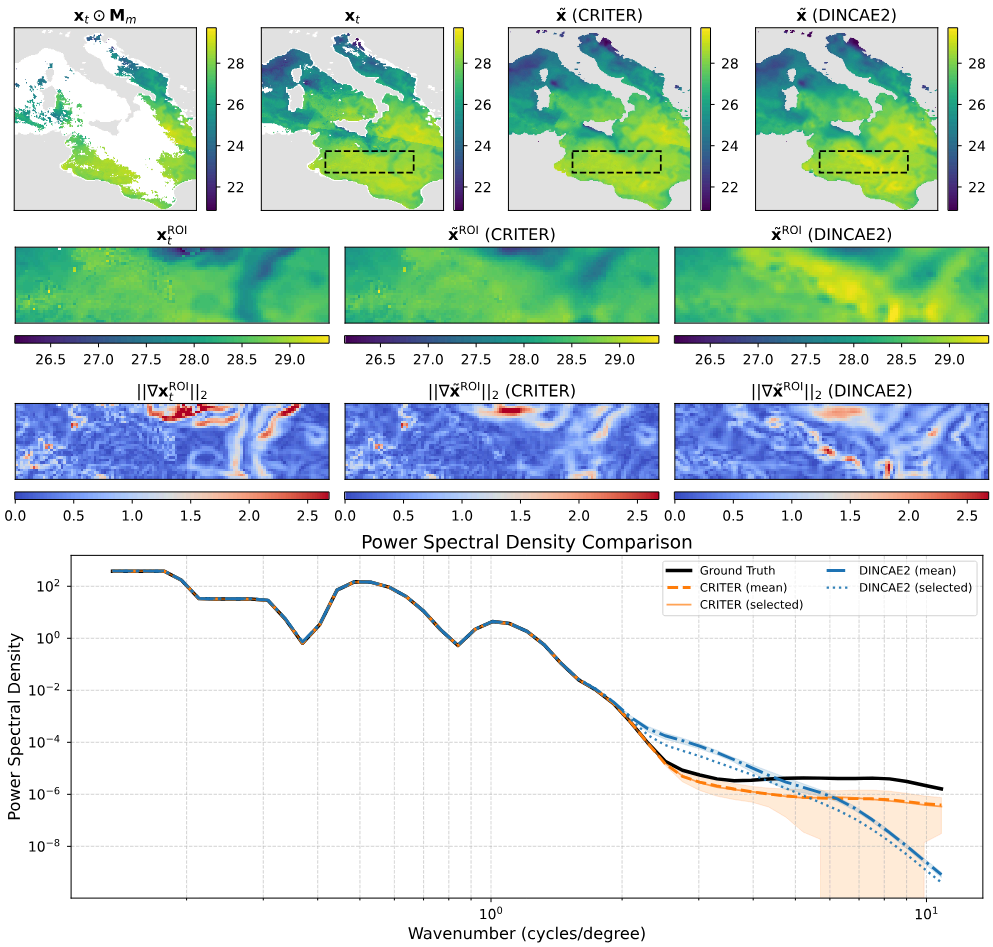

Figure C3. Same as Figure 7, but for a different sample.