Supplementary information to Multiscale neural assimilation scheme for Near-Real-Time SST reconstruction from satellite observations

Maxime Beauchamp,[1,2] Ioanna Karagali[1] , Guisella Gacitúa[1] , Jacob L Høyer[1] , Maxime Ballarotta[3] , Ronan Fablet[2] .

[1] *Danish Meteorological Institute, Sankt Kjelds Plads 11, 2100 Copenhagen, Denmark*

[2] *IMT Atlantique, 655 Av. du Technopôle, 29280 Plouzané, France*

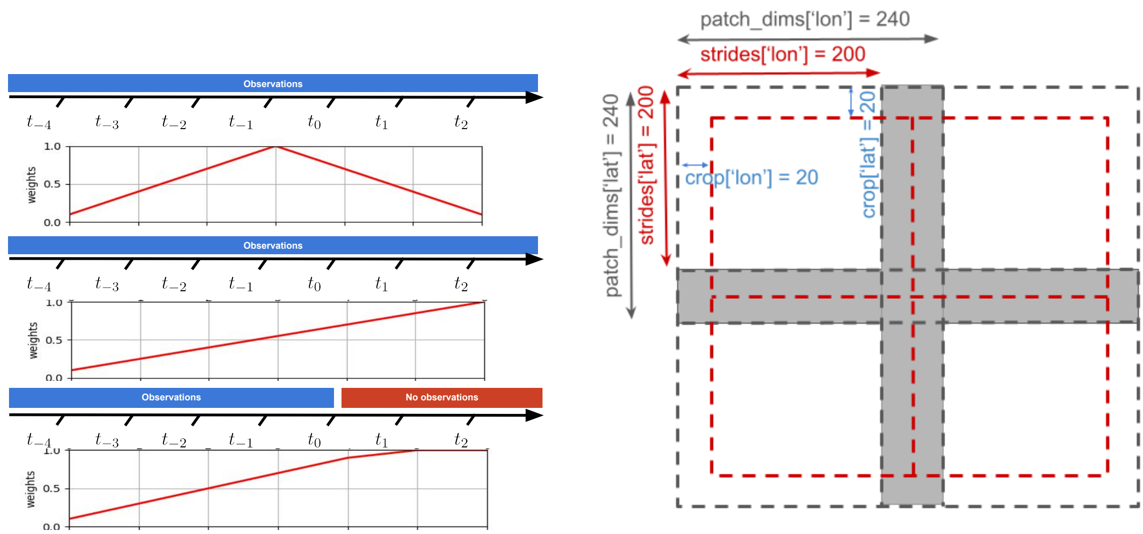
[3] *CLS (Collecte Localisation Satellites), 11 Rue Hermès, 31520 Ramonville-Saint-Agne, France*

Corresponding author: Maxime Beauchamp, maxb@dmi.dk ; maxime.beauchamp@imt-atlantique.fr

1. Complementary details on the training strategy

a. Weighted loss for near-real time product

In the previous applications of the 4DVarNet scheme, the goal was to reconstruct or interpolate a state at the center of a predefined data assimilation window using both past and future in the optimization process. This approach aligns with reanalysis products in oceanography. However, operational products typically rely on short-term forecasts and near-real time data, using only past and, if available, present observations. For learning-based methods, this implies major differences in the optimization strategy because end-users have to identify if the targeted outputs used in the training losses is a specific time in the data assimilation window or the entire window. In terms of training loss function, this means weighting the mean squared error (MSE) loss and the additional regularizers based on the final objective. For interpolation, a triangular-shaped weighting function is typically used, with the highest weights at the center of the assimilation window. For near-real time product, the highest weights at the last time step of the window. Finally, for short-term forecasts (when no observations are available), the approach can be extended by assigning the highest weights to the lead times being predicted. Fig. 1a shows the difference in terms of time weighting within the 7-day data assimilation window for various applications of the 4DVarNet scheme used in this study.



(a) Optimization weights along the data assimilation window

(b) Patches dimension

FIG. 1: (a) Different optimizatiom weights used along the data assimilation window in the outer training loss of the neural scheme depending if we consider a mapping, NRT or short-term forecast inverse problem, and (b) patch-related parameters (dims, strides and crop) of the anomaly field for the high-resolution anomaly field reconstruction

Along this line, because we use a patching strategy with overlapping areas to prevent at most significant discontinuities in the final reconstruction, the MSE loss is also spatially weighted to consider only the center of the patch, cropping the latter based on the combination of patch dimensions and strides used to define the sampling and decomposition of the domain, see Fig. 1b.

b. Training on the single observations

As mentioned before, most of 4DVarNet applications have been carried out in a supervised configuration, where the targeted outputs are fully known. This is related to Observation System Simulated Experiments (OSSE) where model outputs are used as a Ground Truth which makes possible a direct comparison between the outputs of the neural scheme and the targets. But in many situations, such a setting is not available, and only the single real-world observations, together with some potential physical or prior informations conditioning can be used. In this case, a possible extension to provide a satisfactory training configuration consists of removing some observations from the original dataset used as inputs, see Fig. 2. These observations are then used in the training loss to evaluate the neural scheme's ability to fill data gaps. This approach aims to ensure that the reconstruction method generalizes well accross the domain and achieves satisfactory validation performance.

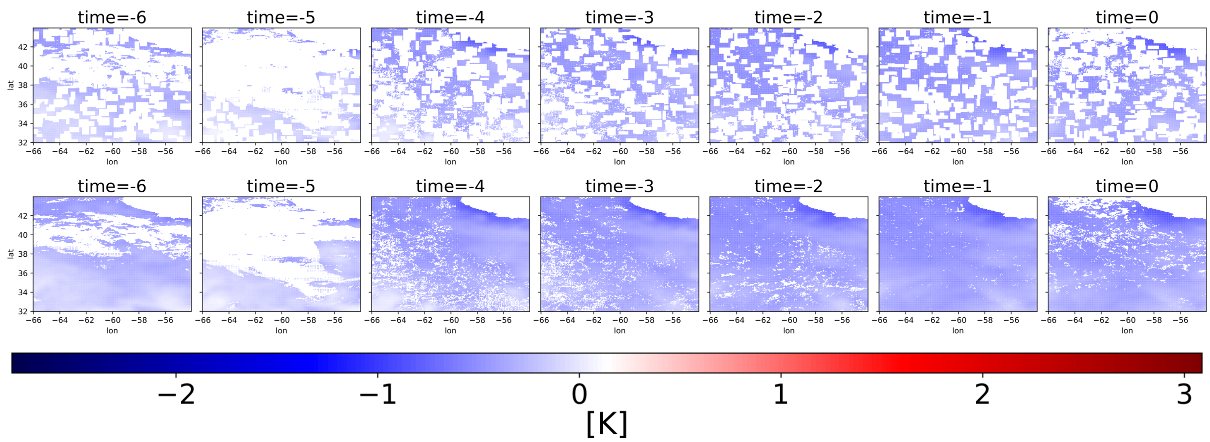


FIG. 2: Example of the self-supervised training strategy used on the observations along the seven days data assimilation window. [Top]: actual inputs used by the neural scheme, random patches (rectangular shapes) are removed from the actual observations. [Bottom]: true observations used to compute the reconstruction loss. For inference after training, the same set of observations will be used to build the NRT reconstruction, corresponding to the last day of the data assimilation window

c. Learning scheme and benchmark strategy

The model is trained using the gappy daily L3S observations from April 2019 to December 2020. During training, random patches of data are removed from the observations so that the reconstruction has to learn from this missing values to generalize on any non-informed areas. The test period corresponds to the entire year 2021. For the benchmarking, DMI-OI and 4DVarNet are applied on a pseudo-L3S dataset in which the SLSTR sensor is removed from the original L3S product and kept aside for validation.

In addition, while the coarse resolution training is unique, we assess the performance of our model on three subdomains $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$ of the whole North and Baltic seas domain. These three domains have already been identified as validation criteria for the DMI-OI release in the CMEMS catalogue (E.U. Copernicus Marine Service Information (CMEMS), Marine Data Store (MDS) 2024). To assess whether a training for the entire domain compete or not with local trainings on these domains, we provide additional results with three additional 4DVarNet anomaly models, one for each validation subdomains.

2. Spectral analysis through space and time

Since we aim for a multiscale implementation of the neural scheme, defining the coarse resolution is essential. We downscale the actual resolution of both L3 observations and L4 interpolated products by a factor of 10 to deal with datasets on $0.2^\circ \times 0.2^\circ$. This considerably reduces the domain size, allowing the entire area to be process at once while accumulating gradients during the forward pass of training. To assess how coarsened products behave in terms of actual spatial resolution retrieved, we provide spectral analyses using Welch method (Welch 1967) to compute power spectrum densities (PSD) from 1-dimensional overlapping segments of data (both observations and reconstruction maps). Because SST obervations are gappy, computed fixed-lengths dimensional segments over two-dimensional maps is challenging and most of the current strategies rely on a regularly-gridded based computations over longitude, latitude and time. As an additional contribution in this paper, we propose an innovative strategy to compute segments of fixed-length anywhere on the domain so that they do not have any missing data due to partial observations or land. To achieve this sampling, we randomly select a large number of locations over the domain, and evaluate the associated segments of the pre-defined length in randomly selected directions. We keep only the segments with no missing data and stop Algorithm 1 when the number of overlapping segments is sufficiently large to evaluate the spectrum (around a thousand). Fig. 3 shows a typical sampling on the entire domain, in which the colorscale relates to the day of the year the segments belongs to. This novel approach to spectrum analysis for SST is particularly useful when applied to large domains, where building segments without post-processed interpolation of missing data is challenging. Notably, the same strategy is applied to the three smaller subdomains, with the length of spatial segments adjusted to fit their actual size. Additional figures in this Supplementary Informations Section 3.b present the seasonal spectral analysis for each subdomain in details.

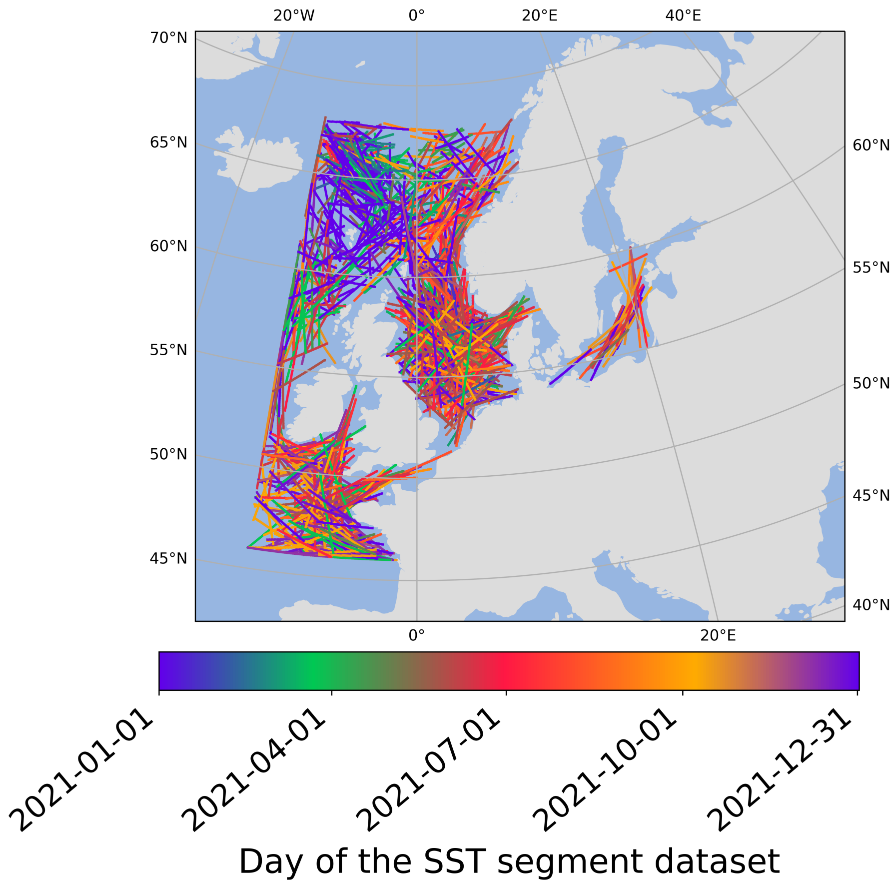


FIG. 3: Daily 400km overlapping segments computed for PSD analysis in 2021. The colorscale relates to which day of the year the segment is collected so that both reconstruction maps and SLSTR observations have no missing data along the segments. It is to note the preferential random sampling depending on the subdomains. It is for instance easier to collect segments in the summer season for the interior Baltic

Algorithm 1 Random Sampling of SST segments

Input: Domain $\mathcal{D}$, predefined segment length L, required number of segments N_{target}
Initialize: Set of valid segments $\mathcal{S} \leftarrow \emptyset$
while $|\mathcal{S}| < N_{\text{target}}$ **do**
    Randomly select a location $x \in \mathcal{D}$
    Randomly select a direction $d \in [0, 360]]$
    Define segment $s \leftarrow$ segment of length L starting from x in direction d
    if s has no missing data **then**
        Add s to the set of valid segments $\mathcal{S}$
    end if
end while
Output: Set of valid segments $\mathcal{S}$

Fig. 4 displays the power spectral density (PSD) as:

- the yearly averaged PSD (2021) (top panel) for coarse (0.2° resolution) NRT reconstructions (both DMI-OI and 4DVarNet, resp. as solid and dashed orange lines) and observations. Black solid line refers to the SLSTR independent sensor ;

- the signal-to-noise ratio (SNR, bottom panel), defined as 1 minus the ratio between the PSD of the error (reconstructions vs observations) and the PSD of the observations. This metric is appropriate to quantify the accuracy of the reconstructions relative to the observations, with higher scores indicating a better match between the two. A cutoff of 0.5 in the PSD SNR is used to determine the yearly averaged minimal spatial scales (in km) identified by the different reconstructions. This threshold is a key aspect in evaluating how well the reconstructions capture the spatial features observed in the data.

DMI-OI reaches at best 77km scales in the SLSTR observational dataset, while 4DVarNet is 16km below. This was expected because contrary to the further anomaly retrieval, we only need a coarse approximation to run the multiscale approach. To simplify the process, we train the coarse version of 4DVarNet in a supervised configuration using a coarse DMI-OI as the target. As a result, the target performs better in the coarse version.

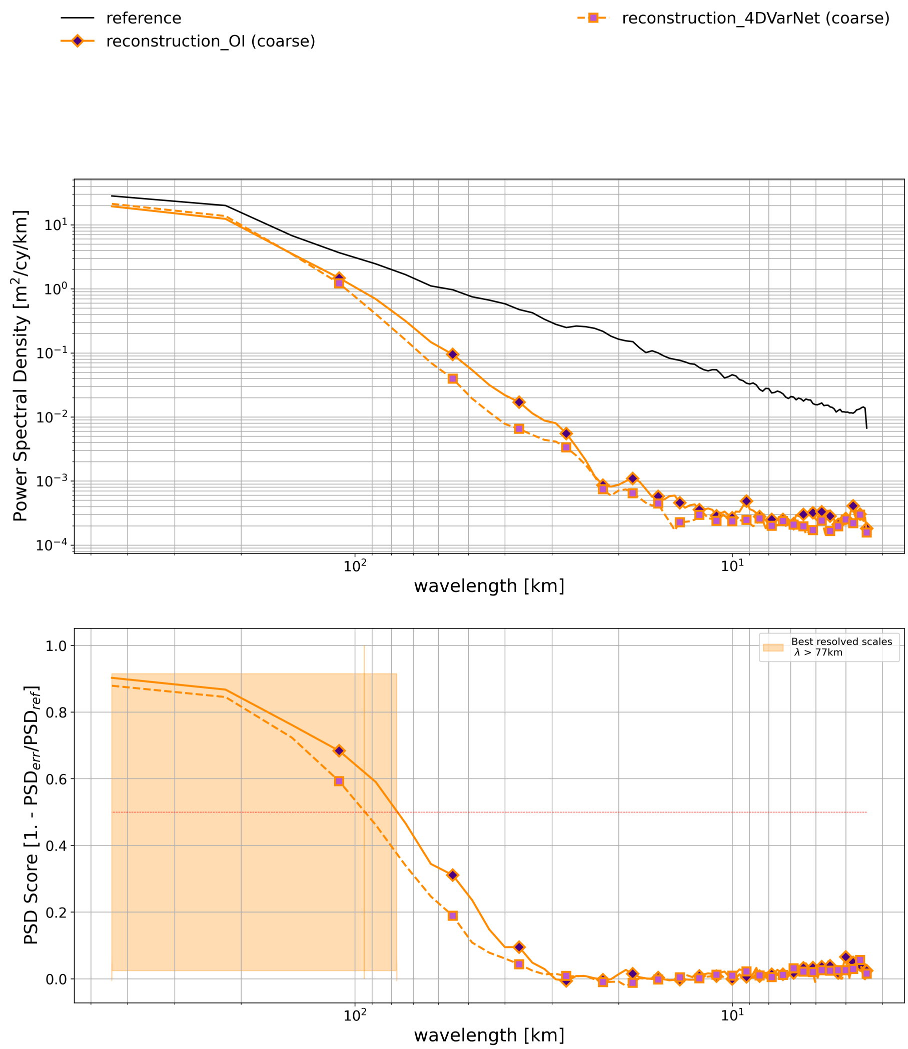


FIG. 4: [Top]: 2021 yearly averaged power spectral density (PSD) for coarse (0.2° resolution) NRT reconstructions and observations. [Bottom] PSD score, i.e. 1 minus the ratio between the PSD of the error (reconstructions-observations) and the PSD of the observations.

3. Complementary validation using SLSTR SST

This Section presents a consolidated assessment of the neural reconstruction scheme, focusing on subdomain-specific performance using SLSTR Sea Surface Temperature (SST) data as a validation reference. The analysis evaluates yearly and seasonal RMSE gains and spectral power density scores for different scales, highlighting how the neural-based approach compares against the operational DMI-Optimal Interpolation (OI) baseline.

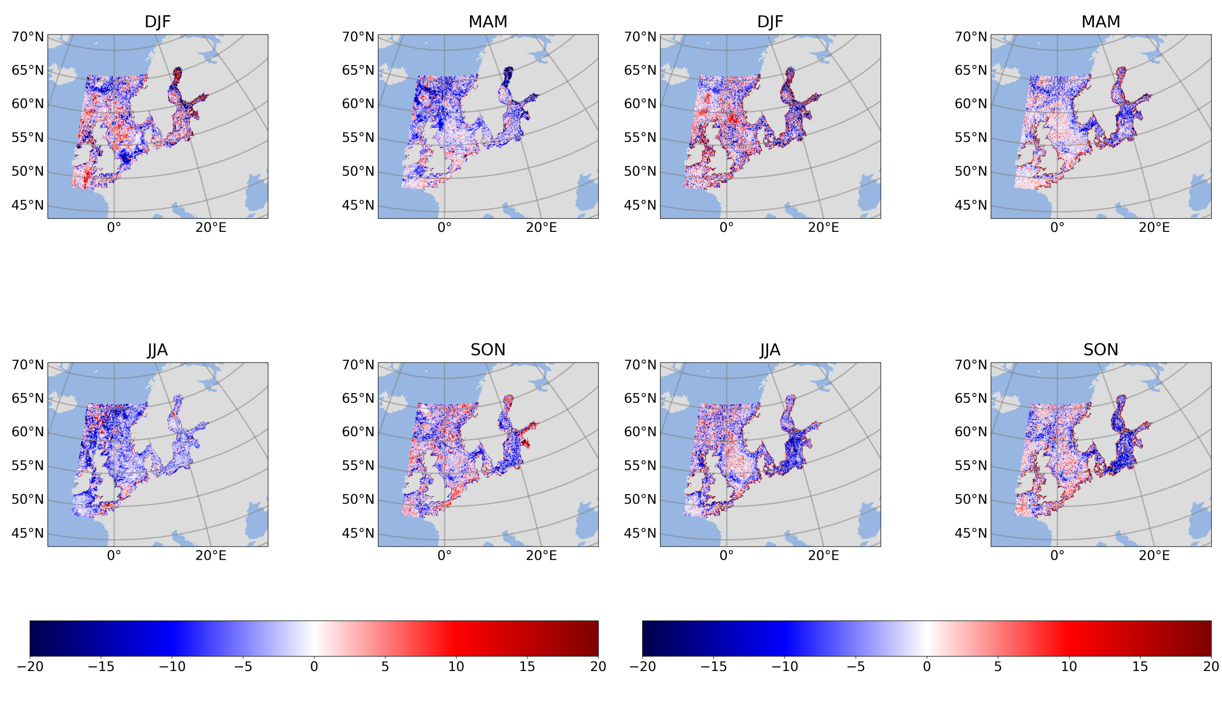
			μ(RMSE)	σ(RMSE)	λ_x (km)
Spring	All Baltic	OI	0.42	0.16	61
		4DVarNet	0.41	0.15	57
	Baltic ($\mathcal{D}_1$)	OI	0.38	0.45	62
		4DVarNet (regional training)	0.38	0.45	48
		4DVarNet ($\mathcal{D}_1$ local training)	0.38	0.45	52
	Baltic ($\mathcal{D}_2$)	OI	0.36	0.13	66
		4DVarNet (regional training)	0.34	0.13	47
		4DVarNet ($\mathcal{D}_2$ local training)	0.35	0.13	43
	Baltic ($\mathcal{D}_3$)	OI	0.37	0.13	NaN
		4DVarNet (regional training)	0.38	0.13	40
		4DVarNet ($\mathcal{D}_3$ local training)	0.32	0.13	32
Summer	All Baltic	OI	0.37	0.13	65
		4DVarNet	0.35	0.13	51
	Baltic ($\mathcal{D}_1$)	OI	0.32	0.17	56
		4DVarNet (regional training)	0.32	0.17	52
		4DVarNet ($\mathcal{D}_1$ local training)	0.32	0.17	45
	Baltic ($\mathcal{D}_2$)	OI	0.37	0.13	66
		4DVarNet (regional training)	0.35	0.13	46
		4DVarNet ($\mathcal{D}_2$ local training)	0.36	0.13	59
	Baltic ($\mathcal{D}_3$)	OI	0.32	0.18	62
		4DVarNet (regional training)	0.32	0.18	32
		4DVarNet ($\mathcal{D}_3$ local training)	0.31	0.17	32
Autumn	All Baltic	OI	0.58	0.17	65
		4DVarNet	0.56	0.16	52
	Baltic ($\mathcal{D}_1$)	OI	0.47	0.19	56
		4DVarNet (regional training)	0.45	0.18	50
		4DVarNet ($\mathcal{D}_1$ local training)	0.45	0.18	47
	Baltic ($\mathcal{D}_2$)	OI	0.53	0.29	65
		4DVarNet (regional training)	0.51	0.29	36
		4DVarNet ($\mathcal{D}_2$ local training)	0.51	0.29	52
	Baltic ($\mathcal{D}_3$)	OI	0.54	0.27	65
		4DVarNet (regional training)	0.52	0.27	39
		4DVarNet ($\mathcal{D}_3$ local training)	0.52	0.27	32
Winter	All Baltic	OI	0.34	0.08	60
		4DVarNet	0.33	0.08	48
	Baltic ($\mathcal{D}_1$)	OI	0.27	0.10	54
		4DVarNet (regional training)	0.27	0.10	52
		4DVarNet ($\mathcal{D}_1$ local training)	0.27	0.10	53
	Baltic ($\mathcal{D}_2$)	OI	0.36	0.19	66
		4DVarNet (regional training)	0.35	0.18	49
		4DVarNet ($\mathcal{D}_2$ local training)	0.35	0.18	49
	Baltic ($\mathcal{D}_3$)	OI	0.36	0.11	NaN
		4DVarNet (regional training)	0.34	0.11	51
		4DVarNet ($\mathcal{D}_3$ local training)	0.34	0.11	40

TABLE 1: Seasonal performance of DMI-OI vs 4DVarNet for the entire North and Baltic Seas domain, and the three subdomains. Seasonally-averaged RMSE (K), seasonal RMSE standard deviation (K), and minimal spatial scale resolved (km). NaN indicates that no spatial scales are correctly retrieved based on the 0.5 signal-to-noise ratio cutoff. This happens sometimes for the DMI-OI interpolation in interior Baltic.

A key aspect is the identification of minimal spatial scales retrieved by each reconstruction scheme, offering insights into the neural model's effectiveness across distinct subregions and timeframes, as indicated by detailed seasonal analysis plots.

Table 1 contains the RMSE-based spectral metrics for every season, domains and scales. This provides a full overview of the current capabilities of both scheme on the test dataset. Multiscale 4DVarNet implementation provides in most of the situations a better or similar agreement with SLSTR validation dataset. Some drops of performance can be observed for the winter season and on the fine scales retrieval, likely related to the training scheme. Further improvements of the actual setup would consist in getting rid of the smoothing kernel of the OI outputs scheme we kept in the neural architecture for fair comparison between the two methods. This smoothing was convenient to remove all the non-physical patterns in the reconstructions due to specific sampling in the SST dataset that can lead to highlight poorly representative observations in the final L4 product. Additional work could help to identify a proper regularization term in the outer loss function to enhance the use of these situations, while keeping the actual underlying high-resolution in the observations.

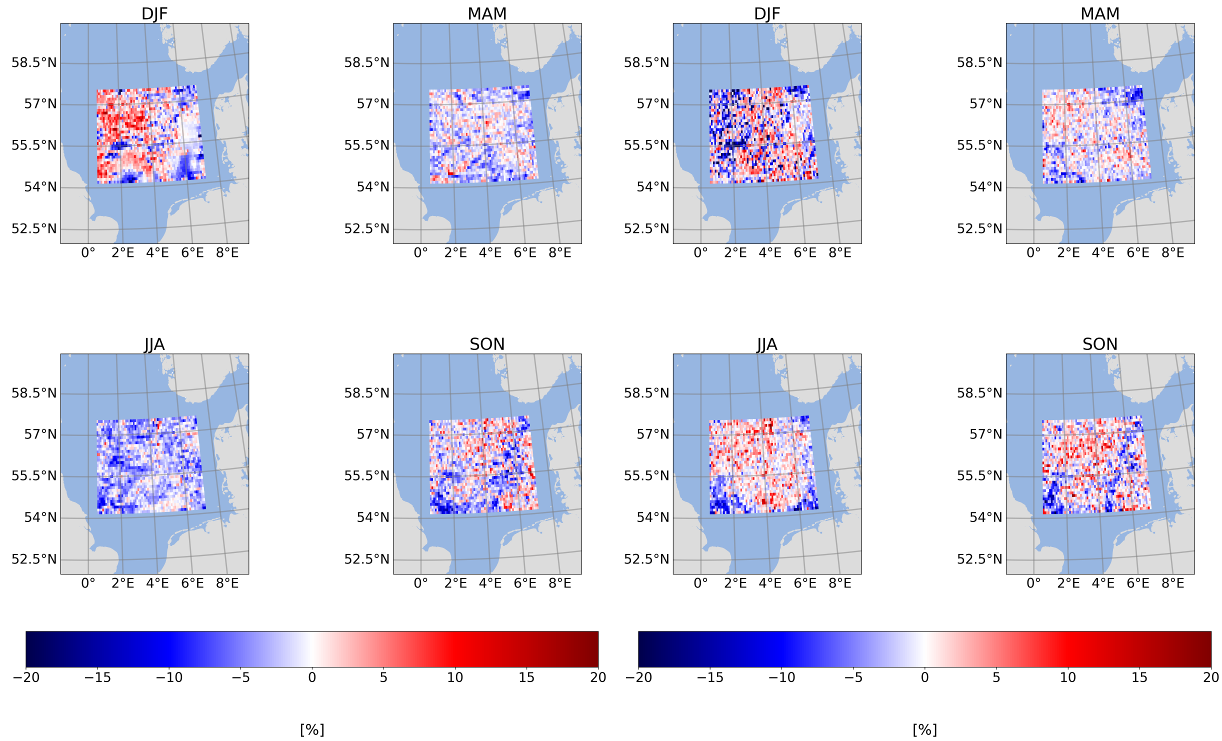
a. Seasonal RMSE improvements



(a) All scales

(b) Scales ≤ 50km

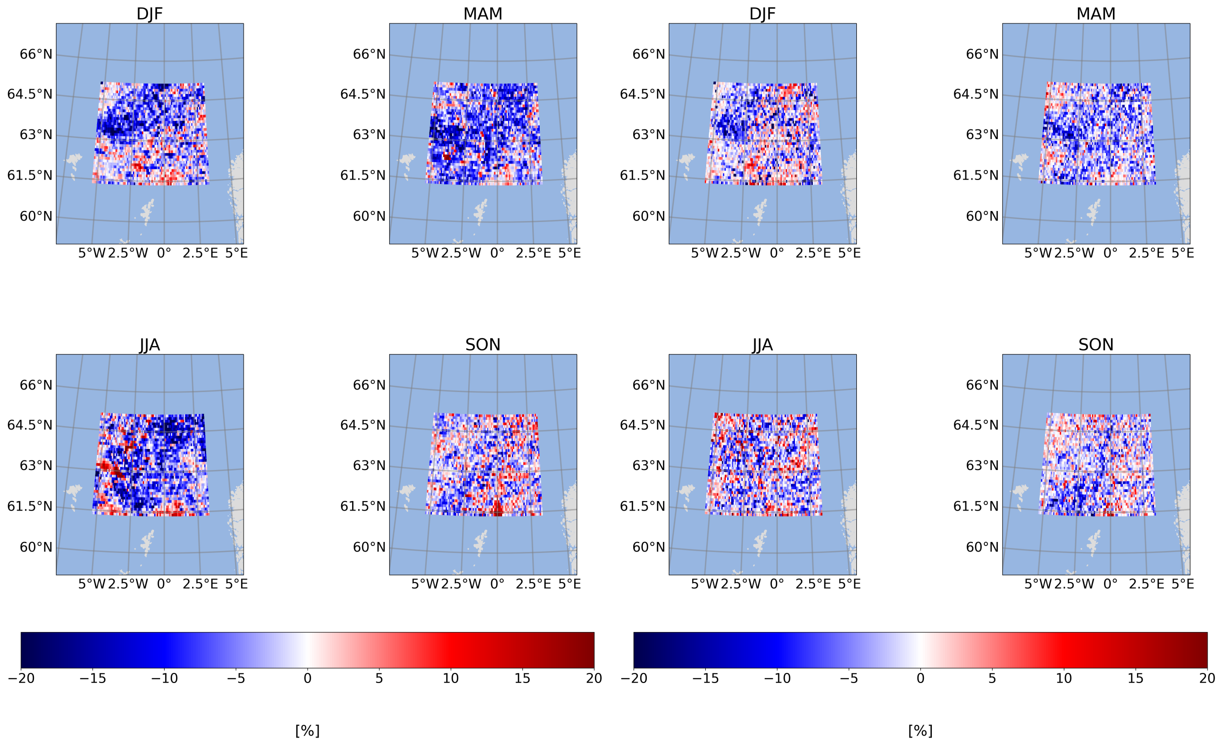
FIG. 5: 2021 seasonal averaged RMSE gain of 4DVarNet vs DMI-OI in bins of .1°×.1°. Blue indicates an average improvement of the neural scheme compared to the baseline. Red indicates a deterioration of the RMSE.



(a) All scales

(b) Scales ≤ 50km

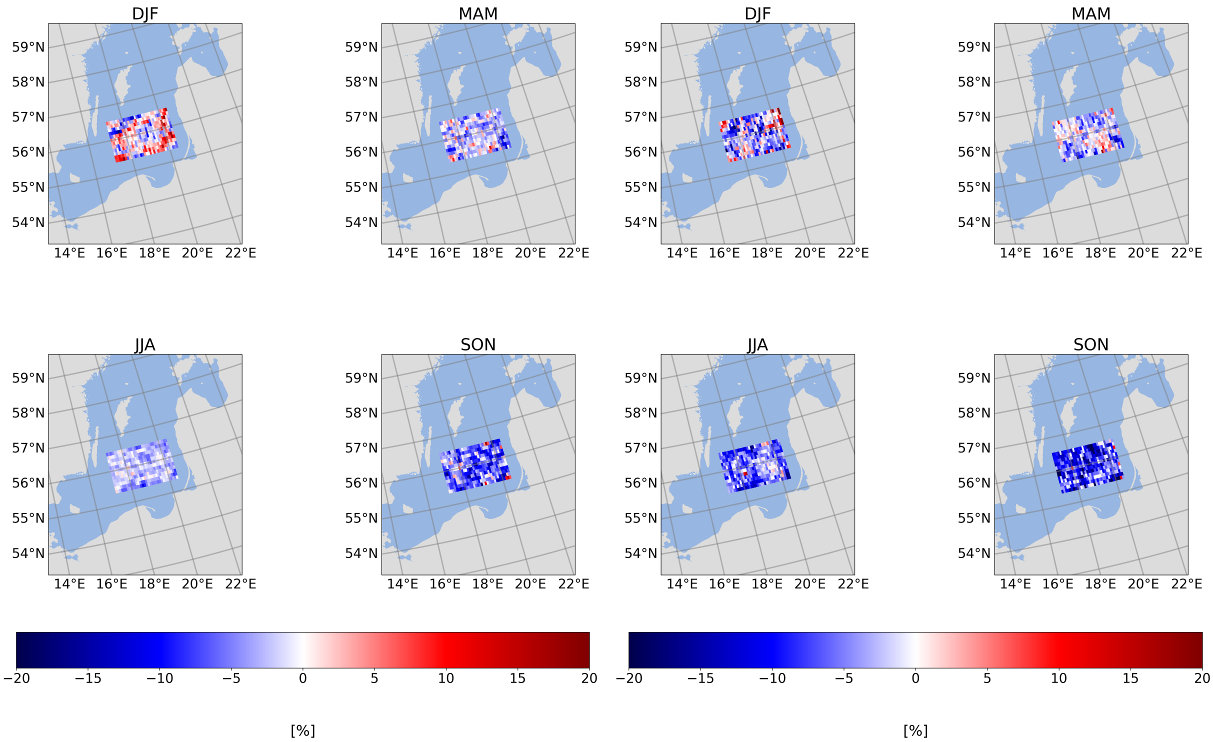
Fig. 6: Domain $\mathcal{D}_1$. 2021 seasonal RMSE gain of 4DVarNet vs DMI-OI in bins of $.1^\circ{\times}.1^\circ$. Blue indicates an average improvement of the neural scheme compared to the baseline ; red indicates a deterioration of the RMSE.



(a) All scales

(b) Scales ≤ 50km

FIG. 7: Domain $\mathcal{D}_2$. 2021 seasonal RMSE gain of 4DVarNet vs DMI-OI in bins of $.1°\times.1°$. Blue indicates an average improvement of the neural scheme compared to the baseline ; red indicates a deterioration of the RMSE.



(a) All scales

(b) Scales ≤ 50km

FIG. 8: Domain $\mathcal{D}_3$. 2021 seasonal RMSE gain of 4DVarNet vs DMI-OI in bins of $.1° \times .1°$. Blue indicates an average improvement of the neural scheme compared to the baseline ; red indicates a deterioration of the RMSE.

While the seasonal analysis may show neighboring pixels with both RMSE gains and degradations, this pattern is largely explained by the aggregation into coarser seasonal bins and the limited density of reference data in some regions, which can locally amplify apparent inconsistencies.

b. Seasonal spectral analysis

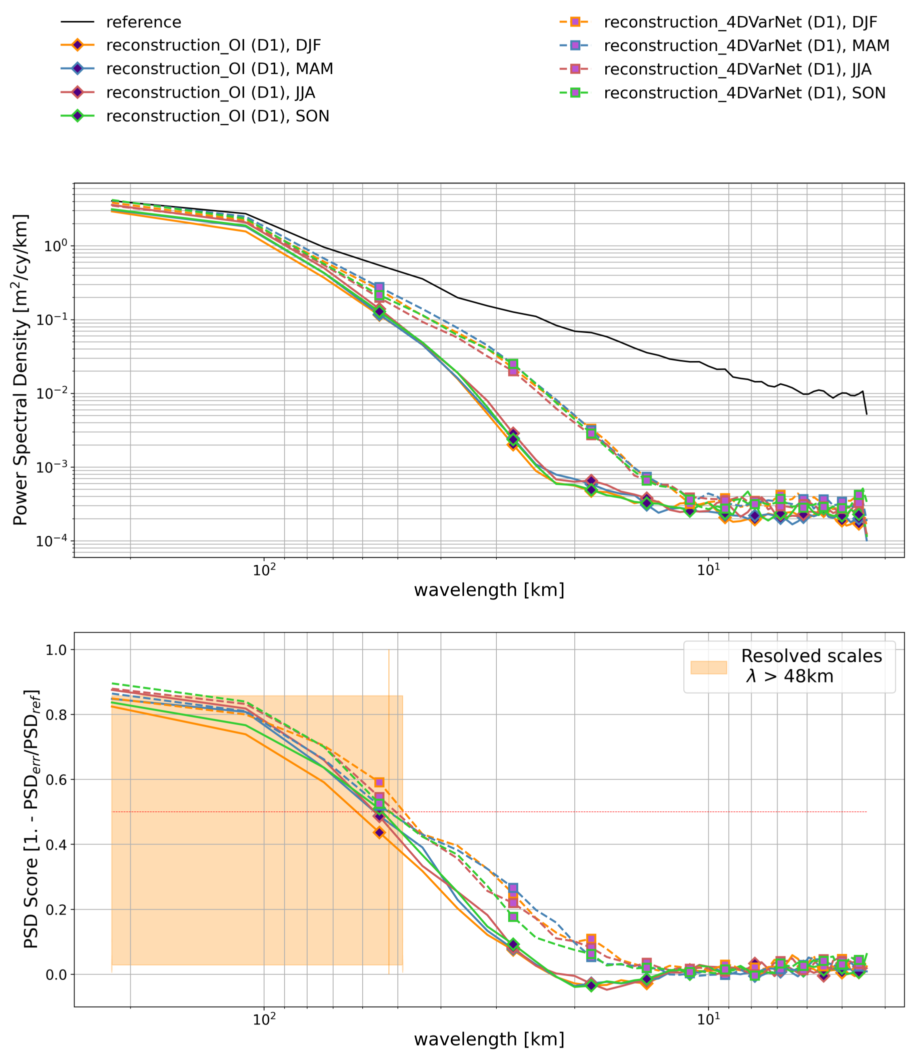


FIG. 9: Domain $\mathcal{D}_1$. [Top]: 2021 seasonal averaged power spectral density (PSD) for NRT reconstructions and observations. [Bottom] PSD score, i.e. 1 minus the ratio between the PSD of the error (reconstructions-observations) and the PSD of the observations. A cut-off of 0.5 for this signal-to-noise ratio in the PSD is used to identify the seasonal average minimal spatial scales (km) retrieved by the different reconstructions.

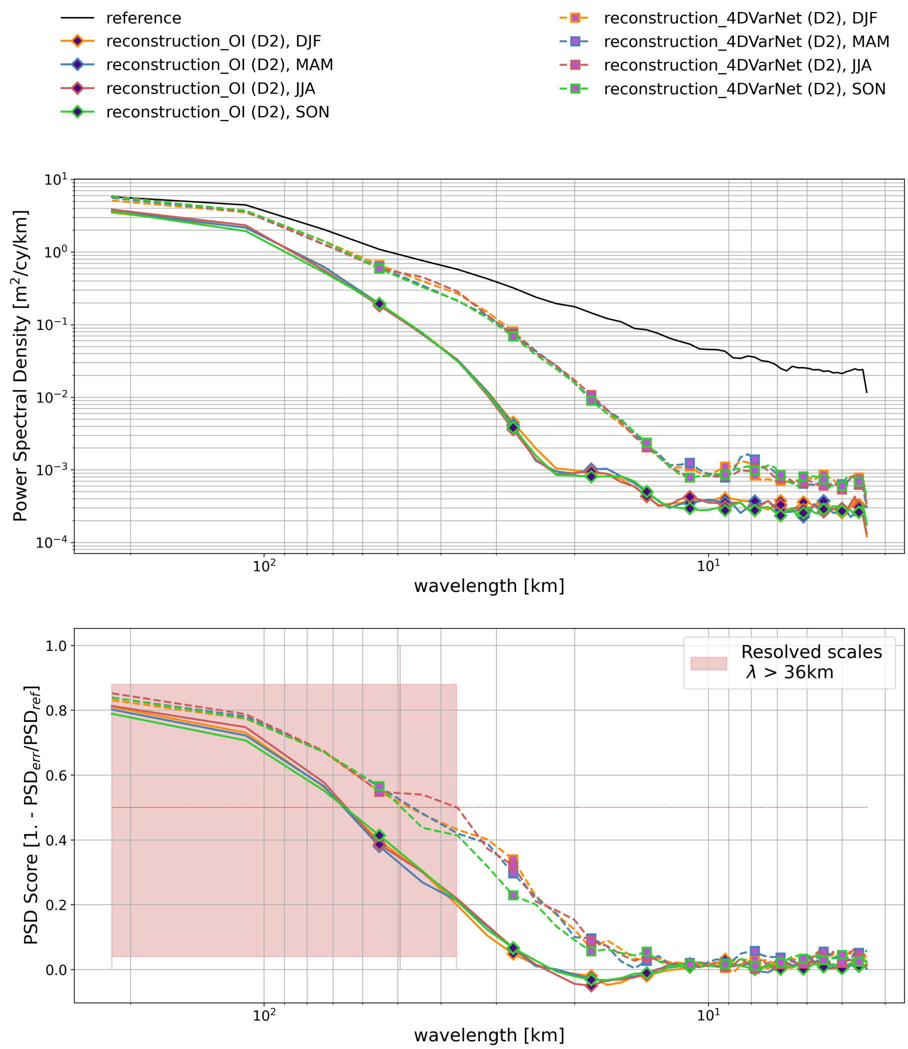


FIG. 10: Domain $\mathcal{D}_2$. [Top]: 2021 seasonal averaged power spectral density (PSD) for NRT reconstructions and observations. [Bottom] PSD score, i.e. 1 minus the ratio between the PSD of the error (reconstructions-observations) and the PSD of the observations. A cut-off of 0.5 for this signal-to-noise ratio in the PSD is used to identify the seasonal average minimal spatial scales (km) retrieved by the different reconstructions. Seasons are identified by the first letters of the months used to compute seasonal metrics, i.e. [DJF] for Winter, [MAM] for Spring, [JJA] for Summer and [SON] for Autumn

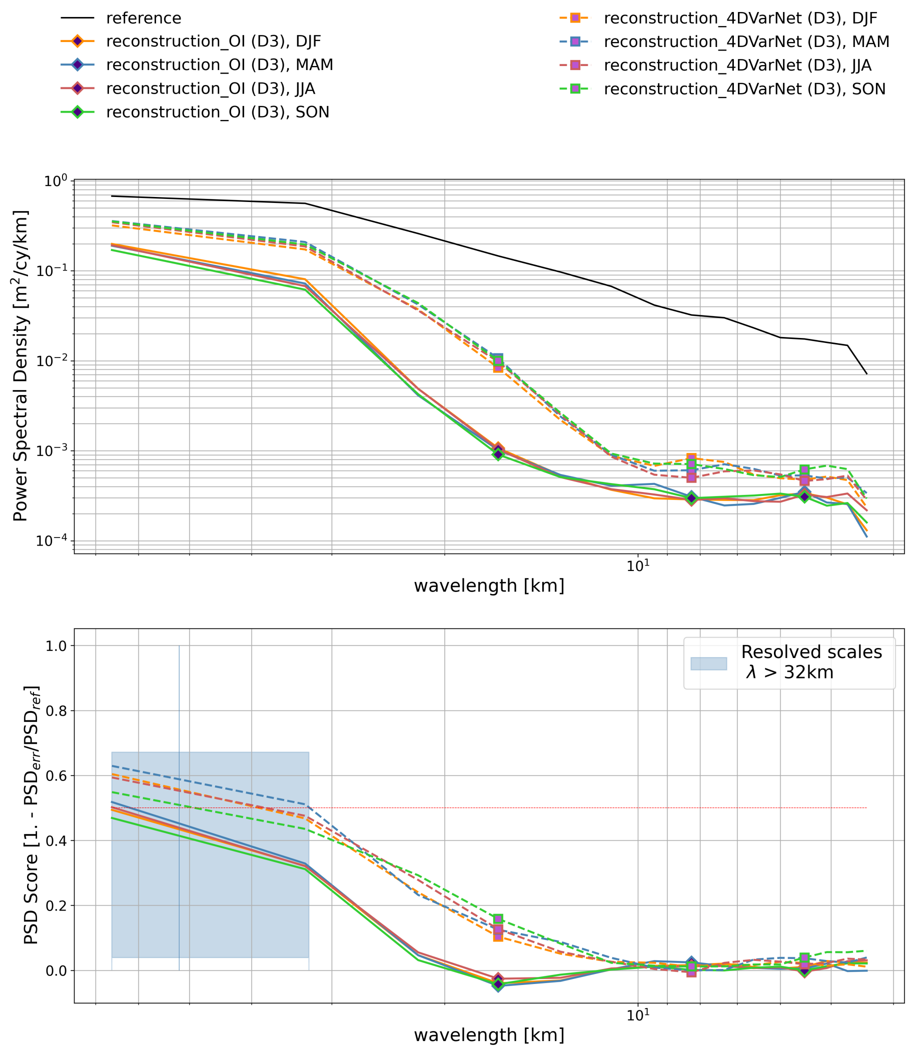


FIG. 11: Domain $\mathcal{D}_3$. [Top]: 2021 seasonal averaged power spectral density (PSD) for NRT reconstructions and observations. [Bottom] PSD score, i.e. 1 minus the ratio between the PSD of the error (reconstructions-observations) and the PSD of the observations. A cut-off of 0.5 for this signal-to-noise ratio in the PSD is used to identify the seasonal average minimal spatial scales (km) retrieved by the different reconstructions. Seasons are identified by the first letters of the months used to compute seasonal metrics, i.e. [DJF] for Winter, [MAM] for Spring, [JJA] for Summer and [SON] for Autumn

4. Complementary validation using in situ observations

Validation statistics between the DMI-OI and the 4DVarNet products against in situ observations not included in the analyses were performed. Table 2 shows the number of points (NUM), mean (μ), standard deviation (σ) and RMSE for each data set collected by: drifting buoys, ship, argo floats and mooring buoys in the North and Baltic Seas region during 2021. Overall, the mean bias

		NUM	μ	σ	RMSE
Drifter	DMI-OI	57426	-0.00	0.48	0.48
	4DVarNet	56497	-0.01	0.49	0.49
Ship	DMI-OI	6978	-0.53	1.31	1.42
	4DVarNet	6840	-0.54	1.35	1.45
Argo	DMI-OI	814	0.17	0.64	0.66
	4DVarNet	805	0.16	0.63	0.65
Mooring	DMI-OI	1899	0.10	0.44	0.46
	4DVarNet	1869	0.09	0.46	0.46

TABLE 2: Validation statistics of biases between the L4 products and in situ datsets for year 2021.

is in good agreement for both products with a maximum difference of ± 0.01. Validation against ship measurements show higher mean bias and standard deviation due to the large discrepancy of depth observation between ship and satellite datasets. The difference in number of matchups for the entire year, in which there are more points for the DMI-OI than for the 4DVarNet, is a result of the current configuration in the 4DVarNet recontruction that excludes the first and last week of the year. However, this doesn't significantly affect the statistical output. Fig. 12 further demonstrate the agreement between the two products assessed.

Fig. 13 shows the monthly mean statistics during the assessed year, 2021. The mean and standard deviation between the two L4 products remain consistent throughout the year, except in comparison with ship data, where persistently high magnitudes are observed, worsening during September-October period. Additionally, a greater discrepancy between the DMI-OI and 4DVarNet biases is observed at the beginning and end of the year, likely due to the lower number of data points considered in the 4DVarNet analyses during these periods.

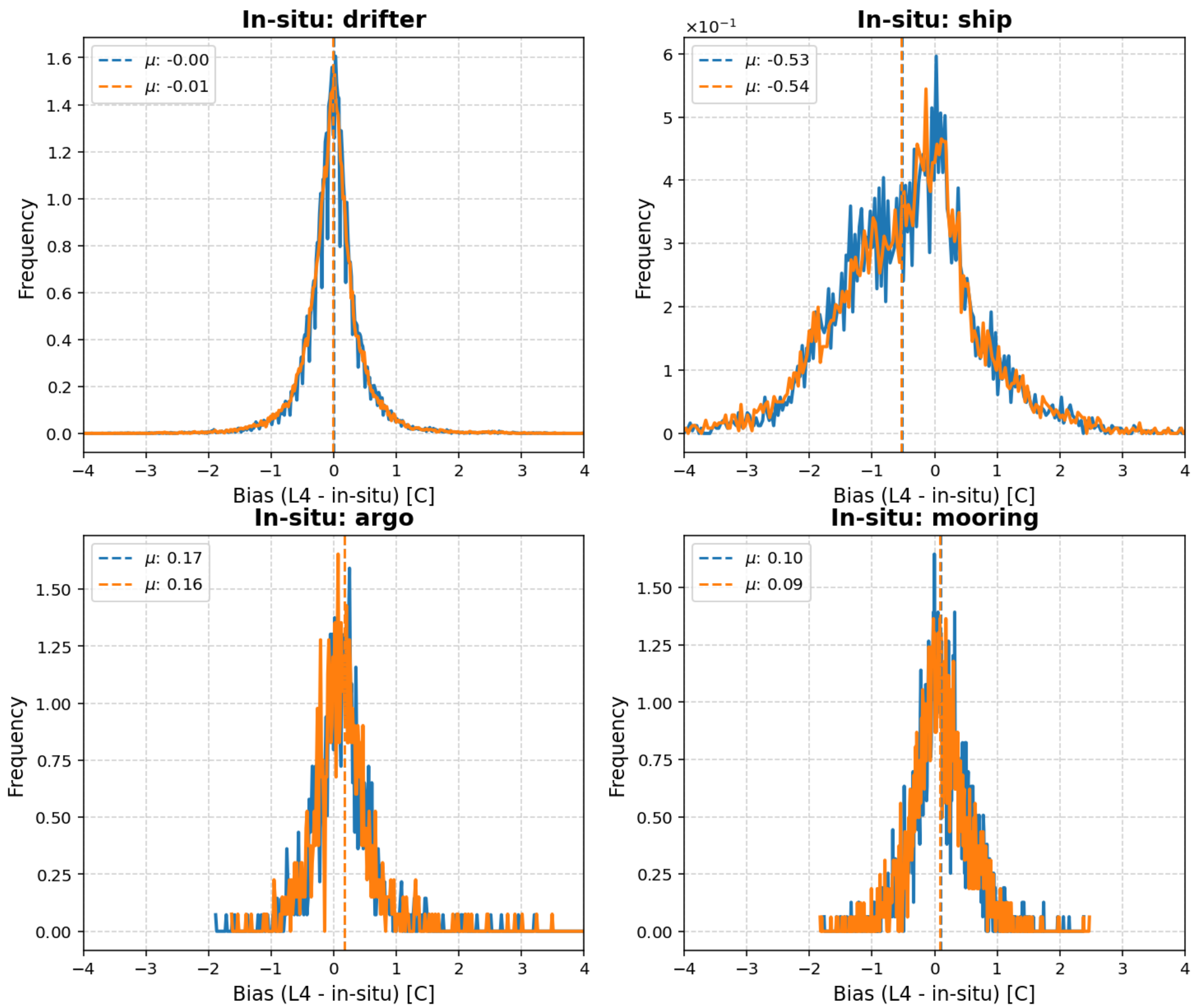


FIG. 12: Comparison of biases between L4 SST output and the in-situ data for the four sensors analised. DMI-OI is depicted in blue and 4DVarNet in orange.

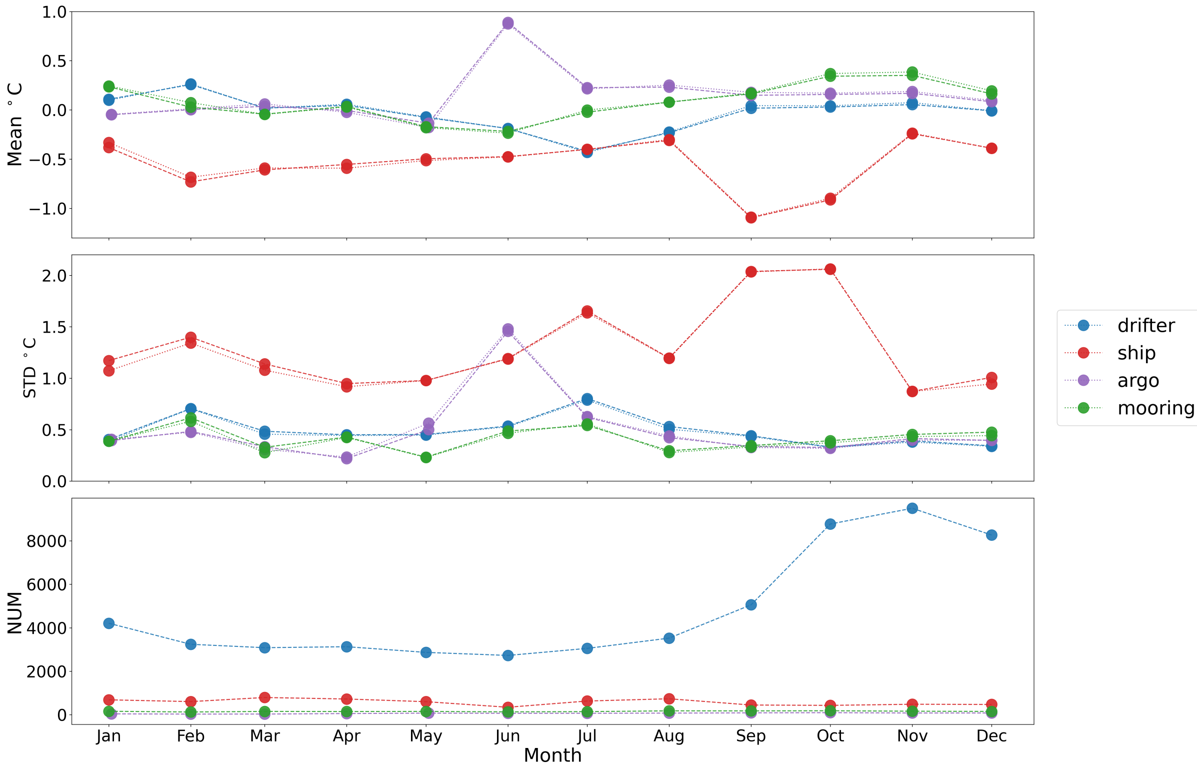


FIG. 13: Monthly mean statistics between the L4 SST output and the in situ data for year 2021. From top to bottom: Mean, STD and Number of matchups. Dotted and dashed lines represent the bias of DMI-OI and 4DVarNet against in situ respectively.

References

E.U. Copernicus Marine Service Information (CMEMS), Marine Data Store (MDS). Baltic Sea Sea Surface Temperature Analysis L4, 2024. URL `https://doi.org/10.48670/moi-00155`.

P. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, June 1967. ISSN 1558-2582. doi: 10.1109/TAU.1967.1161901.