

VAE BASED MULTI-FIELDS NEURAL DATA ASSIMILATION FOR SEA ICE MODEL

Anonymous authors

Paper under double-blind review

ABSTRACT

This study presents a neural data assimilation system based on a variational autoencoder (VAE) for improving sea ice forecasts in high-resolution numerical models. We propose a multi-field assimilation approach that simultaneously processes several physical fields, leveraging a modern VAE architecture enhanced with pixel-wise self-attention mechanisms to capture complex spatial and cross-field correlations. Our method is validated using real-world satellite observations (Sentinel-3 SRAL and AMSR2) and operational data from the NEMO ocean model with integrated sea ice component (SI3). Results demonstrate that the framework effectively assimilates sparse and noisy observations, reducing errors in sea ice concentration estimates and improving forecast accuracy. Crucially, we demonstrate the compatibility of the neural assimilation solution with the NEMO restart mechanism, enabling seamless integration into operational forecasting pipelines. This work bridges the gap between machine learning-based assimilation and practical ocean modeling, offering a scalable, non-Gaussian alternative to traditional methods like 3D-VAR.

1 INTRODUCTION

Significant trend in Earth Sciences is the growing interest in the Arctic region, driven by the rapid decline in sea ice cover (Babb et al., 2023) and the profound environmental, economic, and geopolitical changes this entails (Kortsch et al., 2015; Shu et al., 2023; Dvoynikov et al., 2021). Numerical models in the Arctic and Antarctic regions are notoriously harder to run and calibrate (Pan et al., 2023; Allende et al., 2024). Sea ice forecasting requires considering the entangled relationships between wind, ocean physics, and plastic deformations inside the ice itself to be relevant. These forecasts hold practical value for ice-impeded navigation in addition to scientific value.

Naturally, Earth Sciences move towards increasing the resolution of climate models (Moreno-Chamarro et al., 2025; Olason et al., 2021; Selivanova et al., 2024) and expanding the volume of accumulated observational data (Copernicus Climate Change Service (C3S); National Aeronautics and Space Administration) in order to improve the prediction of the extreme events and circulations. This trend is driven by advances in computational power, improved satellite and sensor technologies.

Numerical models are a cornerstone in ocean simulation, incorporating modern knowledge about the complex physical processes that govern ocean dynamics. Because of the chaos, inherent to all such models, small uncertainty of the initial conditions grows to unpredictable model behavior in a few days horizons of simulation. Data assimilation is a necessary tool to condition these numerical models on observations and improve their quality. Classical data assimilation algorithms rely on the assumptions of linear model dynamics and Gaussian noise. However, as the resolution of the model increases, these assumptions become less valid, which requires a departure from these constraints to better capture the complexities of high-resolution systems (Carrassi et al., 2018).

The data assimilation task can be formulated as the process of iterative updating of the physical fields of the model x_b (background) based on the observed data y . The target is to get closer to an unknown true state of the fields x_a (analysis) (Bannister, 2008a;b). The data assimilation process must account for the spatial and temporal relationship of the physical fields.

There are several approaches to data assimilation. Non-neural approaches will be called in our text classical data assimilation methods. In most cases, it comes down to filtering techniques such as

Best Linear Unbiased Estimator (BLUE) and Kalman Filter (KF) (Kalman, 1960; Jazwinski, 1970; Evensen, 1994; 2003) or to the optimization of a cost function as 3 and 4-Dimensional Variational (3D-VAR, 4D-VAR) (Bannister, 2008a;b; Courtier et al., 1998; Rabier et al., 2000) data assimilation. Neural networks have shown a remarkable ability to learn complex and non-linear relationships between physical fields in oceanography (Zhao et al., 2024). There are several studies that combine traditional data assimilation algorithms and modern neural networks (Blanke et al., 2024; Arcucci et al., 2021; Penny et al., 2022; Cai et al., 2024; Tian, 2024; Hatfield et al., 2021; Mack et al., 2020; Peyron et al., 2021; Farchi et al., 2021; Barthelemy et al., 2022; Melinc & Zaplotnik, 2024).

Most existing data assimilation models focus mainly on simplified systems and benchmark examples, such as the Lorenz-63 or Lorenz-96 systems (Blanke et al., 2024; Arcucci et al., 2021; Penny et al., 2022; Tian, 2024; Peyron et al., 2021; Farchi et al., 2021). These toy models serve as foundational testbeds for evaluating the performance of assimilation techniques due to their ability to capture essential features of chaotic dynamics while remaining computationally tractable. However, their simplicity often limits the direct applicability of these methods to more complex, real-world systems characterized by high dimensionality and intricate spatial correlations.

Observation data can be sparse and contain errors and inaccuracies. The true values of the physical field are unknown, as are the true values of the model and the observation error. We should use some estimation for calculation and validation. This is especially noticeable when calculating the matrix B . Classical data assimilation algorithms operate under Gaussian assumptions. For example, the ice concentration field has a significantly non-Gaussian error distribution, while the Ensemble Kalman Filter EnKF was used to assimilate this field (Lisæter et al., 2003).

1.1 3D-VAR AND 4D-VAR

This is family of iterative data assimilation algorithms that are based on cost function optimization. Cost function for 3D-VAR is

$$J(x) = (x - x_b)^T B^{-1} (x - x_b) + (y - Hx)^T R^{-1} (y - Hx) \quad (1)$$

where B is a background error covariance matrix, R is an observation error covariance matrix, and H is the forward operator. For 4D-VAR the model operator M is added:

$$J(x_0) = (x_0 - x_b)^T B^{-1} (x_0 - x_b) + \sum_{i=1}^N (y - H(x_i))^T R^{-1} (y - H(x_i)) \quad (2)$$

where $x_0 = x(t_0)$ are the physical fields at the beginning of the assimilation window and $x_i = x(t_i) = (\prod_{k=1}^i M_k) x(t_0)$. It can be shown that Kalman Filter minimizes the same function (Brasseur, 2006).

1.2 ARTIFICIAL NEURAL NETWORKS FOR DATA ASSIMILATION

An alternative to classical data assimilation schemes is the integration of artificial neural networks (ANNs) to replace or enhance components of traditional systems. There are two main strategies for incorporating ANNs into data assimilation. The first involves using a neural network to distill classical methods (Arcucci et al., 2021; Farchi et al., 2021; Zavala-Romero et al., 2025), such as 3D-VAR, 4D-VAR, and ETKF, capturing their key features in a more computationally efficient form. The second approach integrates ANNs directly into the assimilation process by replacing specific components of traditional algorithms (Penny et al., 2022; Tian, 2024; Hatfield et al., 2021; Mack et al., 2020; Peyron et al., 2021; Barthelemy et al., 2022; Melinc & Zaplotnik, 2024), potentially improving adaptability and providing more accurate error estimates. The second approach is of greater interest because it enables a more flexible and adaptive data assimilation process. By replacing key components of traditional algorithms with ANNs, we can take advantage of their ability to learn complex non-linear relationships within the data. This has the potential to improve accuracy, reduce computational costs, and enhance robustness in dynamically changing ice conditions.

Classical data assimilation algorithms face two major challenges. The first is the use of the covariance matrix, which can be extremely large and may fail to capture nonlinear interactions between variables. The second is the dynamical operator, which is either computationally expensive or based on overly simplistic approximations.

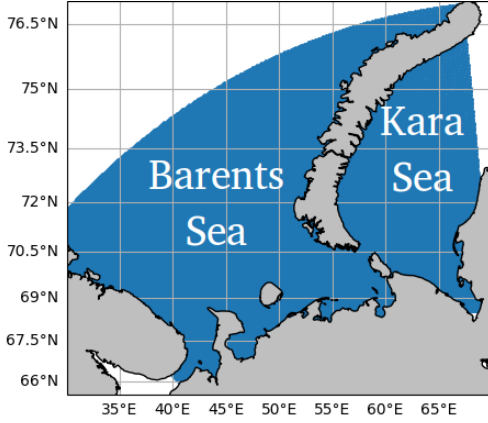


Figure 1: Area of interest for this work.

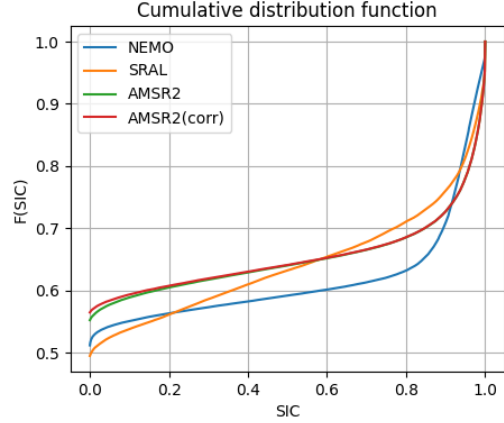


Figure 2: Cumulative distribution functions (CDFs) of sea ice concentration from different data sources.

The variational autoencoder (VAE) is widely used as a replacement or complement to the covariance matrix, enabling dimensionality reduction while preserving complex variable interactions. VAEs impose constraints on data distribution, making them particularly suitable for physical data fields with nonlinear dependencies. In (Mack et al., 2020), a VAE is applied within the 3D-Var algorithm to reduce the dimensionality of physical fields while capturing intervariable relationships, tested on pollution data from Elephant and Castle, London. In (Peyron et al., 2021), the ensemble Kalman filter is applied in the latent space of an autoencoder and tested on the Lorenz-96 model. In (Melinc & Zaplotnik, 2024), an autoencoder completely replaces the covariance matrix in the 3D-Var algorithm. The approach is tested on temperature data at the 850-hPa pressure level from the ERA5 reanalysis. The latter study is conceptually similar to our work but considers only a single physical field and does not assess the possibility of restarting the model with the assimilated field.

This paper makes the following contributions:

1. We propose a new data assimilation algorithm that works with multiple geophysical fields by leveraging VAE architecture with self-attention layers in the latent space.
2. We test our algorithm on different setups and show that it outperforms the baselines: classical 3D-VAR and a VAE-based approach (Melinc & Zaplotnik, 2024).
3. We integrated our algorithm inside the operational forecasting ocean model Nucleus for European Modeling of the Ocean (NEMO) to assimilate real satellite observations. We have shown that neural network based data assimilation improves the forecasts quality.

2 DATA

The area of interest for this work is the Barents Sea and the Kara Sea (see Figure 1). The Kara and Barents Seas are not covered with multi-year ice de Gelis et al. (2021). For the selected region of the Kara Sea, the freeze-up begins rapidly in December, with melting in April. The selected region of the Barents Sea has a completely open sea throughout most of the year.

3 MODEL

3.1 BACKGROUND

As an assimilation background we use ocean data that are generated by our production numerical forecasting system that consists of numerical Weather Research and Forecasting model (WRF) and Nucleus for European Modeling of the Ocean (NEMO) of version 4.0 with integrated Sea Ice Modeling Integrated Initiative (SI3). Spatial resolution of the modeling dataset is around 3-4 km, we

used ORCA12 grid and Drakkar configuration that is considered state of the art for high resolution ocean modeling. Since NEMO requires atmosphere forcing we also used these data as additional features (these data are generated by the atmosphere model Weather Research and Forecasting WRF of version 4.4). For the period of interest, the dataset contains daily prediction of the next 72 hours for ocean and atmosphere variables. Data range from 2015 to 2023. In this study, we focus on three key fields: the sea ice concentration field (where data assimilation is performed), the sea ice thickness field, and the sea ice temperature field. Data from 2015 to 2021 (inclusive) are used to compute statistical properties and train the variational autoencoder (VAE). 2022 and 2023 are reserved for validating the accuracy of field reconstruction and conducting data assimilation experiments.

3.2 OBSERVATIONS

The following observational data was considered:

1. Sentinel-3 Altimetry (SRAL) Sea Ice Thematic Product Aublanc et al. (2025). The dataset includes satellite tracks with an along-track resolution of up to 330 meters. The variable "sea_ice_concentration_20_ku" provides ice concentration data and is available as an auxiliary product. The primary data source is the OSI-430-a product, which has a 2-day latency and a spatial resolution of 25 km. Therefore, greater scientific interest lies in the variable "surf_type_class_20_ku", which characterizes surface type (open ocean, floes, lead, unclassified) based on sea ice concentration and waveform peakiness, and the variable "freeboard_20_ku" (radar freeboard). In future product versions, the latter is planned to be converted into ice thickness.
2. ASI AMSR2 Spreen et al. (2008) sea ice concentration data. These data have a spatial resolution of 6.25 km and are available in near-real time (NRT) with a latency about 25 hours. A key advantage of AMSR2 is its all-weather, day-and-night operational capability, enabled by passive microwave measurements. However, a known limitation is the reduced accuracy in sea ice concentration retrieval during melt seasons due to changes in surface emissivity caused by snow and ice melt Pang et al. (2018).

We have determined that the most promising problem to address is the assimilation of data tracks, especially since Sentinel-3 is expected to enhance its thematic product on sea ice with ice thickness information. For model-to-model assimilation, we use sea ice concentration data from a different model year, selected along SRAL tracks. For satellite-to-model assimilation, we employ AMSR2 sea ice concentration data along SRAL tracks, adjusted using the surf_type_class_20_ku flag: values are set to zero for the "open ocean" class, while for "floes" and "lead" classes with zero concentration, the nearest non-zero value is assigned.

3.3 VALIDATION

For model-to-model assimilation, the model fields corresponding to the dates of the assimilation tracks are used for validation and comparison. For satellite-to-model assimilation, AMSR2 data serve as the validation data set.

3.4 DATA ANALYSIS

To compare sea ice concentration data from different sources along the Sentinel-3 track between December 2019 and April 2020, their cumulative distribution functions (CDF) should be analyzed (Fig. 2). Data were collected for identical dates and identical spatial domains in all sources. The NEMO model (blue line) exhibits a lower probability of zero ice concentration compared to AMSR2 (green line), indicating systematic overestimation of ice cover in the model output. The SRAL-derived 'sea_ice_concentration_20_ku' variable (orange line) shows markedly different slope characteristics than both NEMO and AMSR2. This divergence stems from the coarse resolution of OSI-430-a product (source of this variable), which introduces spatial averaging effects that increase the number of intermediate values.

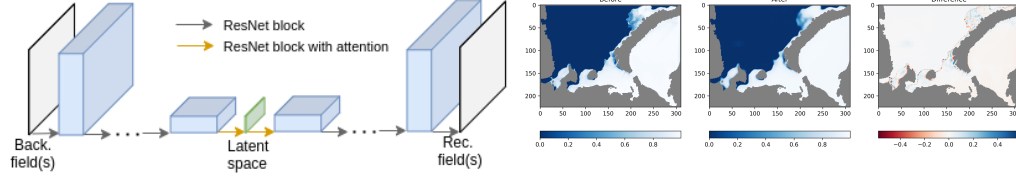


Figure 3: Variational Autoencoder architecture. Figure 4: Example of VAE data reconstruction.

4 MODEL

In the 3D-VAR algorithm, spatial relationships between the components of a field, as well as inter-field correlations, are represented in the background error covariance matrix. In this study, we propose using a variational autoencoder (Kingma & Welling, 2014) to capture and represent these patterns. A related approach was previously applied to temperature data at the 850-hPa pressure level from the ERA5 reanalysis, as reported in (Melinc & Zaplotnik, 2024). We used the VAE architecture from (Melinc & Zaplotnik, 2024) as our baseline, denoting it as *base_vae_1f*. In our work, we employ a different autoencoder architecture and perform experiments using both single-field and multi-field datasets. The architecture we propose is inspired by the VAE architectures widely used in stable diffusion models Rombach et al. (2022), which have demonstrated strong generative and representation learning capabilities. The main characteristics of our architecture 3 are as follows: (1) the latent space is implemented as a feature map rather than a vector, (2) ResNet-based blocks are used throughout the encoder and decoder, and (3) attention mechanisms are introduced in the middle layers to enhance feature interaction.

We conducted experiments with a date-conditioned autoencoder where: dates were converted to day-of-year (DOY) values. DOY was encoded as a cyclic variable using sine/cosine transformation: $(\sin(2\pi \cdot \text{DOY}/366), \cos(2\pi \cdot \text{DOY}/366))$ which are transformed by the linear layer and concatenated with in the latent space.

4.1 TRAINING

The variational autoencoder learns to reconstruct fields using the reparameterization trick. Mean squared error (MSE) is used as the reconstruction loss, while the KL divergence is computed for the Gaussian case (Kingma & Welling, 2014). We also experimented with the addition of a discriminator and SSIM loss, but it did not produce significant improvements.

We used the Lion optimizer (Chen et al., 2023) for training, as it demonstrated a more stable convergence in our experiments. Missing data (over land) are filled with physically adequate values.

4.2 ASSIMILATION

For the baseline data assimilation experiment, we employed a 3D-VAR scheme *3d_var*. The background error covariance matrix B was modeled using a fifth-order quasi-Gaussian function (Gaspari & Cohn, 1999) with a length scale of 100 km.

The latent space assimilation algorithm (Algorithm 1) is presented below. The key feature of the algorithm: assimilation is performed in the autoencoder’s latent space. The latent space field is adjusted via backpropagation to ensure the reconstructed (analysis) field better approximates the observational data.

It is important to discuss the features of the *Loss* function. The loss function consists of three terms: (1) the observation error, (2) the background field error, and (3) the latent-space error. The second and third terms serve as regularizations to prevent a significant deviation from the model’s original field and are assigned smaller weighting coefficients.

$$\begin{aligned} \text{Loss}(x_a, y, x_b, z, z_0) = & w_y \text{MSE}(H(x_a), y) \\ & + w_b \text{MSE}(x_a, x_b) + w_z \text{MSE}(z, z_0) \end{aligned} \quad (3)$$

Algorithm 1 Latent Space Assimilation (LSA)**Require:** *Encoder* and *Decoder* - part of pretrained VAE, *Loss* - error function**Input:** x_b - background fields, y - observations**Output:** x_a - analysis fieldsEncode background: $z_0 \leftarrow \text{Encoder}(x_b)$ Initialize optimizable latent: $z \leftarrow z_0$ (requires gradient)**for** iteration $i=0$ to N **do** Decode current latent: $x_a \leftarrow \text{Decoder}(z)$ Compute loss: $L_{total} \leftarrow \text{Loss}(x_a, y, x_b, z, z_0)$ Compute gradient: $\nabla_z L_{total}$

Update latent: optimization step

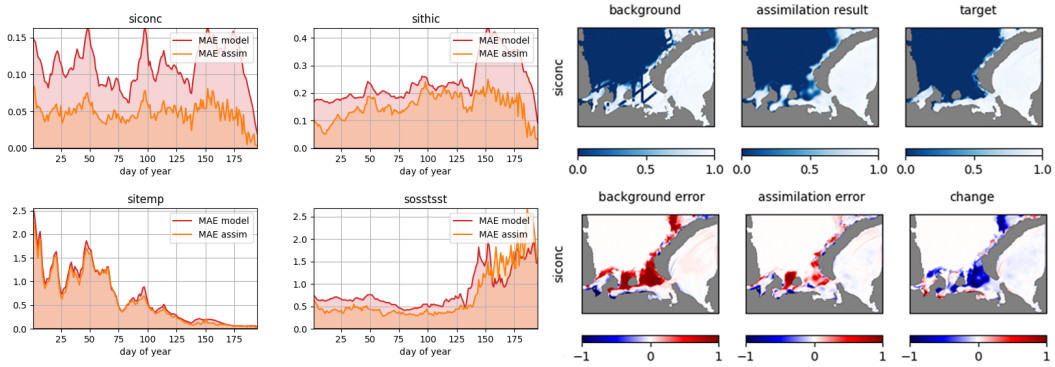
end for **return** $x_a \leftarrow \text{Decoder}(z)$ 

Figure 5: MAE metric values for *vae_4f* model from different physical fields by day of the year on the left. Sea Ice Concentration assimilation example for 12.04 using *vae_4f* model on the right.

Given the slow evolution of ice concentration (typically $< 1\%$ daily change), we tested assimilation of the prior three days data.

5 EXPERIMENTS

Since we aimed at improving the quality of our forecasting system in the real world through sentinel 3 SRAL assimilation, we organized our framework as follows. First, we evaluated the quality of the VAE reconstruction. Secondly, we experimented with model-to-model assimilation to tune out probable overfitting and deal with general architecture search. After that we tested the best resulting models on real-world data and tested our assimilation model inside NEMO forecasting pipeline.

The VAE should reconstruct the fields reasonably well, though perfect accuracy is not the goal, as VAEs inherently balance reconstruction fidelity against the dimensionality of the compressed representation (i.e., latent space sparsity). The goal of the model-to-model assimilation is to estimate the quality of assimilation in neutral setting. Since observation data and model data are noisy and usually distributed differently, thus introducing bias in the quality of assimilation. Model-to-model, on the other hand, demonstrates the quality of assimilation in the physics-driven process that is represented by the numerical model inside NEMO. For the forecasting pipeline, let's call this assimilation scheme production data assimilation.

5.1 RECONSTRUCTION ERROR

The model was trained on data from 2015 to 2021 and 2022 was used for validation. The following fields were utilized: siconc (sea ice concentration), sithic (sea ice thickness), sitemp (sea

ice temperature) and sosstsst (sea surface temperature). An example of sea ice concentration field reconstruction is shown in Figure 4. The results are presented in Table 1 .

The evaluation metrics used are MAE (Mean Absolute Error) and MSE (Mean Squared Error). MAE estimates the average error magnitude, while MSE penalizes larger errors more heavily and smaller errors less severely due to its quadratic nature. The results are presented in Table 1. Features of the architecture and naming of models: **_f** - shows how many physical fields are fed to the model input. **_d*** the latent space is transformed by a linear layer from a feature map into a vector with the specified number of features. **_m** the sitemp field is replaced by the sosstsst field, **_emb** - a condition for the day of the year is added, **_c*** - the number of feature maps in the latent space, 1 by default.

Table 1: VAE reconstruction quality assessment

Model name	siconc MAE	sithic MAE	sitemp MAE	sosstsst MAE
base_vae_1f	0.028 ± 0.002	-	-	-
vae_1f	0.008 ± 0.001	-	-	-
vae_1f_d512	0.023 ± 0.002	-	-	-
vae_3f	0.023 ± 0.002	0.083 ± 0.006	0.094 ± 0.007	-
vae_3f_emb	0.026 ± 0.002	0.090 ± 0.007	0.093 ± 0.008	-
vae_3f_m	0.015 ± 0.001	0.024 ± 0.002	-	0.252 ± 0.006
vae_3f_m_emb	0.016 ± 0.001	0.026 ± 0.002	-	0.254 ± 0.007
vae_4f	0.024 ± 0.002	0.083 ± 0.006	0.085 ± 0.007	0.265 ± 0.008
vae_4f_emb	0.032 ± 0.002	0.099 ± 0.008	0.083 ± 0.007	0.267 ± 0.008
vae_4f_c2	0.018 ± 0.001	0.029 ± 0.002	0.062 ± 0.005	0.210 ± 0.005
vae_4f_c2_emb	0.017 ± 0.001	0.031 ± 0.002	0.061 ± 0.005	0.186 ± 0.003

5.2 QUALITY ESTIMATION

5.2.1 MODEL TO MODEL ASSIMILATION

In this experiment, our objective was to analyze the ability of a trained VAE to capture relationships between values and physical fields. Unfortunately, we do not know the true field values at each time step: The model contains computational errors and depends on the initial initialization, which also introduces uncertainties. Satellite data have measurement errors and interpretation challenges.

Therefore, we conclude that it is necessary to verify the assimilation of the model data into the model data. We sample see ice concentration data along the SRAL tracks from the NEMO forecast for next year $\tilde{y} = x_{i+365}[mask]$ and assimilate them into current data x_i . To evaluate performance, we compare the post-assimilation fields with the prediction of the model x_{i+365} (see Algorithm 2). Since we treat NEMO numerical modeling as physics-based, we expect that a higher quality model will assimilate data closer to the NEMO output.

Algorithm 2 Model-to-model (M2M) data assimilation

Require: $\{x_1, x_2, \dots, x_n\}$ - NEMO simulation fields, $\{y_1, y_2, \dots, y_n\}$ - AMSR corrected by SRAL data

for day in assimilation cycle **do**

 Create mask: $mask = y_i$ is not None

 Sample observation: $\tilde{y}_i = x_{i+365}[mask]$

 Make assimilation: $x_i^a = \text{LSA}(x_i, \tilde{y}_i)$

 Make validation: x_i and x_i^a vs x_{i+365}

end for

Examples of model-in-model assimilation for $3d_var$ vae_4f are illustrated in figures in 7. Only sea ice concentration values are assimilated. It can be seen that vae_4f not only yields a lower error but also produces a sharper ice-water boundary, while $3d_var$ tends to smooth it. The vae_4f results show that when the concentration of ice decreases, the thickness of the ice in the same region also

Table 2: Model to model assimilation result: average MAE

Model name	siconc	sithic	sitemp	sosstsst
background	0.112 ± 0.002	0.242 ± 0.005	0.67 ± 0.04	0.86 ± 0.03
3d_var	0.052 ± 0.001	-	-	-
base_vae_1f	0.080 ± 0.002	-	-	-
vae_1f	0.051 ± 0.001	-	-	-
vae_1f_d512	0.053 ± 0.001	-	-	-
vae_3f	0.060 ± 0.001	0.172 ± 0.004	0.63 ± 0.04	-
vae_3f_emb	0.057 ± 0.001	0.167 ± 0.004	0.64 ± 0.04	-
vae_3f_m	0.061 ± 0.001	0.200 ± 0.004	-	0.84 ± 0.05
vae_3f_m_emb	0.062 ± 0.001	0.213 ± 0.004	-	0.811 ± 0.045
vae_4f	0.0481 ± 0.0009	0.152 ± 0.003	0.615 ± 0.036	0.732 ± 0.039
vae_4f_emb	0.0493 ± 0.0009	0.158 ± 0.004	0.608 ± 0.035	0.663 ± 0.031
vae_4f_c2	0.0508 ± 0.0010	0.168 ± 0.003	0.616 ± 0.035	0.739 ± 0.034
vae_4f_c2_emb	0.0504 ± 0.0009	0.176 ± 0.003	0.608 ± 0.035	0.724 ± 0.037

decreases, while both the temperature of the ice and the temperature of the ocean increase. These patterns align with the target fields and maintain physical consistency with the model. This clearly shows that the VAE successfully captures not only relationships within individual physical variables but also the cross-correlations between different physical parameters.

Figure 5 shows the MSE and MAE metric values for data assimilation using the *vae_4f* model, applied to the fields of ice concentration, ice thickness, ice temperature and upper-layer water temperature over different days of the year. The data is interrupted in mid-July when the ice in the Barents and Kara Seas completely melts. It is evident that the ice concentration and thickness fields are well-correlated, whereas the ice and water temperature fields show weaker dependence on the ice concentration. Furthermore, Table 2 present the MAE metrics for assimilation across different models. We observe that the incorporation of temperature fields improves the accuracy of the ice concentration and thickness assimilation. The *vae_4f* model was selected as the primary because of its best performance in capturing the relationship between ice concentration and thickness.

5.2.2 SAT TO MODEL ASSIMILATION

The next step was data assimilation of AMSR2 ice concentration data corrected by SRAL surface type. We evaluated the quality of assimilation against the AMSR2 data and corrected AMSR2 data that had not yet been assimilated. The assimilation scheme works as follows.

Algorithm 3 Satellite-to-model (S2M) data assimilation

Require: $\{x_1, x_2, \dots, x_n\}$ - NEMO simulation fields, $\{y_1, y_2, \dots, y_n\}$ - AMSR corrected by SRAL data, $\{v_1, v_2, \dots, v_n\}$ - AMSR data.
for day in assimilation cycle **do**
 Make assimilation: $x_i^a = \text{LSA}(x_i, y_i)$
 Make validation: x_i and x_i^a vs v_i and y_{i+1}
end for

The assimilation results for the SRAL-corrected AMSR2 sea ice concentration data are presented in the Table 3. Although the *vae_3f_emb* model showed slightly better metrics, we selected the *vae_4f* model for the final experiment. This decision is based on its superior performance in capturing ice thickness variations in the M2M experiment, coupled with the fact that the difference in results for this experiment is within the margin of error. Samples of assimilation results are in Appendix A.3.

5.3 PRACTICAL APPLICATION

In the last experiment, we integrate the results of the satellite data assimilation into the NEMO pipeline using the restart mechanism. The model restart-ice fields were modified using the *siconc* (sea ice concentration) and *sithic* (sea ice thickness) fields, which assimilated satellite data via the

Table 3: Satellite to model assimilation result

Model name	AMSR2		AMSR2 corrected (track)	
	MSE	MAE	MSE	MAE
background	0.027 ± 0.002	0.049 ± 0.002	0.028 ± 0.002	0.050 ± 0.003
3d_var	0.013 ± 0.001	0.037 ± 0.002	0.019 ± 0.002	0.044 ± 0.002
base_vae_1f	0.019 ± 0.001	0.045 ± 0.002	0.021 ± 0.002	0.051 ± 0.003
vae_1f	0.014 ± 0.001	0.034 ± 0.002	0.019 ± 0.002	0.040 ± 0.002
vae_1f_d512	0.012 ± 0.001	0.033 ± 0.002	0.017 ± 0.002	0.041 ± 0.002
vae_3f	0.012 ± 0.001	0.033 ± 0.002	0.017 ± 0.002	0.040 ± 0.002
vae_3f_emb	0.012 ± 0.001	0.032 ± 0.002	0.017 ± 0.002	0.039 ± 0.002
vae_3f_m	0.014 ± 0.001	0.034 ± 0.002	0.018 ± 0.002	0.043 ± 0.002
vae_3f_m_emb	0.013 ± 0.001	0.036 ± 0.002	0.018 ± 0.002	0.044 ± 0.002
vae_4f	0.013 ± 0.001	0.033 ± 0.002	0.018 ± 0.002	0.041 ± 0.002
vae_4f_emb	0.012 ± 0.001	0.039 ± 0.002	0.018 ± 0.002	0.046 ± 0.002
vae_4f_c2	0.014 ± 0.001	0.033 ± 0.002	0.019 ± 0.002	0.040 ± 0.002
vae_4f_c2_emb	0.013 ± 0.001	0.033 ± 0.002	0.019 ± 0.002	0.041 ± 0.002

vae_4f model. The *sitemp* (ice surface temperature) and *sosstsst* (sea surface temperature) fields were deliberately excluded, as they demonstrated a weaker correlation with the assimilated values. The restart data is modified as told in appendix A.1.

The experiment was organized as follows. A specific date was selected. For this date, data assimilation was performed and the restart file was modified. Subsequently, a 5-day forecast was run using the NEMO model. The forecast results were then compared with the corresponding AMSR2 data for those days. The outcome of this comparison is presented in the Figure 8. Metrics for the experiments are in Table 4.

The experiment demonstrated the feasibility of using fields processed by neural network-based data assimilation within the NEMO computational model for forecasting. The predictions generated by this approach are physically consistent and show good agreement with satellite observations.

Table 4: Application result: MAE with AMSR2 (20-02-2025)

Experiment	1 day	2 day	3 day	4 day	5 day
model	0.142	0.123	0.081	0.084	0.081
model+assimilation	0.079	0.086	0.063	0.072	0.072

6 CONCLUSION

This work is devoted to multi-field data assimilation by neural networks as an alternative to classical approaches. For this we used a variational autoencoder with self-attention layers that helped us to account to cross-correlation between the physics fields. We have shown that such an approach surpasses in terms of quality both classical approach 3D-VAR and the competing approaches from literature based on similar ideas.

Also we tested our approach on the real operative sea ice forecasting workflow in Arctic based on Nucleus for European Modelling of the Ocean (NEMO) state-of-the-art ocean modeling framework. We have shown that assimilation of real satellite data in the model quickly improves the model quality, so our approach has strong practical applications.

Possible direction of our approach improvement would be to correct not only a single snapshot but a whole time series in order to assure physical consistency of the forecasts dynamics. The may challenge here lies in learning a good evolution operator for the whole bunch of participating physical fields. In order to improve quality of forecasts even further it would make sense not only to assimilate sea ice data but also to correct atmosphere forcing and increase domain of simulations to capture more oceanic and atmospheric processes in the Arctic.

REFERENCES

- S. Allende, A. M. Treguier, C. Lique, C. B. Montégut, F. Massonnet, T. Fichefet, and A. Barthelemy. Impact of ocean vertical-mixing parameterization on arctic sea ice and upper-ocean properties using the nemo-si3 model. *Geoscientific Model Development*, 17(20):7445–7466, 2024. doi: <https://doi.org/10.5194/gmd-17-7445-2024>.
- R. Arcucci, J. Zhu, S. Hu, and Y. Guo. Deep data assimilation: Integrating deep learning with data assimilation. *Applied Sciences*, 11(3), 2021. doi: <https://doi.org/10.3390/app11031114>.
- J. Aublanc, J. Renou, F. Piras, K. Nielsen, S. K. Rose, S. B. Simonsen, S. Fleury, S. Hendricks, N. Taburet, G. D’Apice, A. Chamayou, P. Femenias, F. Catapano, and M. Restano. Sentinel-3 altimetry thematic products for hydrology, sea ice and land ice. *Scientific Data*, 12(714), 2025. doi: <https://doi.org/10.1038/s41597-025-04956-3>.
- D. Babb, R. Galley, S. Kirillov, J. Landy, S. Howell, J. Stroeve, W. Meier, J. Ehn, and D. Barber. The stepwise reduction of multiyear sea ice area in the arctic ocean since 1980. *JGR Oceans*, 128(10), 2023. doi: <https://doi.org/10.1029/2023JC020157>.
- R. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. i: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134(637):1951–1970, 2008a. doi: <https://doi.org/10.1002/qj.339>.
- R. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. ii: Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society*, 134(637):1971–1996, 2008b. doi: <https://doi.org/10.1002/qj.339>.
- S. Barthelemy, J. Brajard, L. Bertino, and F. Counillon. Super-resolution data assimilation. *Ocean Dynamics*, 72:661–678, 2022. doi: <https://doi.org/10.1007/s10236-022-01523-x>.
- M. Blanke, R. Fablet, and M. Lelarge. Neural incremental data assimilation. *arXiv:2406.15076*, 2024. URL <https://arxiv.org/abs/2406.15076>. Accessed: 2025-03-19.
- P. Brasseur. Ocean data assimilation using sequential methods based on the kalman filter. In *Ocean Weather Forecasting: An Integrated View of Oceanography*, pp. 271–316. Springer Science+Business Media B.V., 2006. doi: <https://doi.org/10.1007/1-4020-4028-8>.
- S. Cai, F. Fang, and Y. Wang. Advancing neural network-based data assimilation for large-scale spatiotemporal systems with sparse observations. *Physics of Fluids*, 36(9), 2024. doi: <https://doi.org/10.1063/5.0228384>.
- A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5), 2018. doi: <https://doi.org/10.1002/wcc.535>.
- X. Chen, C. Liang, D. Huang, E. Real, K. Wang, Y. Liu, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu, and Q. V. Le. Symbolic discovery of optimization algorithms. 2023. doi: <https://doi.org/10.48550/arXiv.2302.06675>.
- Copernicus Climate Change Service (C3S). Copernicus climate data store. [Online]. URL <https://cds.climate.copernicus.eu/>. Accessed: 2025-03-19.
- P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ecmwf implementation of three-dimensional variational assimilation (3d-var). i: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998. doi: <https://doi.org/10.1002/qj.49712455002>.
- I. de Gelis, A. Colin, and N. Longepe. Prediction of categorized sea ice concentration from sentinel-1 sar images based on a fully convolutional network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5831–5841, 2021. doi: 10.1109/JSTARS.2021.3074068.
- M. Dvoynikov, G. Buslaev, A. Kunshin, D. Sidorov, A. Kraslawski, and M. Budovskaya. New concepts of hydrogen production and storage in arctic region. *Resources*, 10(3), 2021. doi: <https://doi.org/10.3390/resources10010003>.

- G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(C5):10143–10162, 1994. doi: <https://doi.org/10.1029/94JC00572>.
- G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003. doi: <https://doi.org/10.1007/s10236-003-0036-9>.
- A. Farchi, M. Bocquet, P. Laloyaux, M. Bonavita, and Q. Malartic. A comparison of combined data assimilation and machine learning methods for offline and online model error correction. *Journal of Computational Science*, 55, 2021. doi: <https://doi.org/10.1016/j.jocs.2021.101468>.
- G. Gaspari and S. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554), 1999. doi: <https://doi.org/10.1002/qj.49712555417>.
- S. Hatfield, M. Chantry, P. Dueben, P. Lopez, A. Geer, and T. Palmer. Building tangent-linear and adjoint models for data assimilation with neural networks. *Journal of Advances in Modeling Earth Systems*, 13(9):1–16, 2021. doi: <https://doi.org/10.1029/2021MS002521>.
- H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York and London, 1970.
- R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960. doi: <https://doi.org/10.1115/1.3662552>.
- D. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014. doi: <https://doi.org/10.48550/arXiv.1312.6114>.
- S. Kortsch, R. Primicerio, M. Fossheim, A. Dolgov, and M. Aschan. Climate change alters the structure of arctic marine food webs due to poleward shifts of boreal generalists. *Proceedings of the Royal Society B*, 282(1814), 2015. doi: <https://doi.org/10.1098/rspb.2015.1546>.
- K. A. Lisæter, J. Rosanova, and G. Evensen. Assimilation of ice concentration in a coupled ice–ocean model, using the ensemble kalman filter. *Ocean Dynamics*, 53:368–388, 2003. doi: <https://doi.org/10.1007/s10236-003-0049-4>.
- J. Mack, R. Arcucci, M. Molina-Solana, and Y. Guo. Attention-based convolutional autoencoders for 3d-variational data assimilation. *Computer Methods in Applied Mechanics and Engineering*, 372:1–33, 2020. doi: <https://doi.org/10.1016/j.cma.2020.113291>.
- B. Melinc and Z. Zaplotnik. 3d-var data assimilation using a variational autoencoder. *Quarterly Journal of the Royal Meteorological Society*, 150(761):2273–2295, 2024. doi: <https://doi.org/10.1002/qj.4708>.
- E. Moreno-Chamarro, T. Arsouze, M. Acosta, P.-A. Bretonniere, M. Castrillo, E. Ferrer, A. Frigola, D. Kuznetsova, E. Martin-Martinez, P. Ortega, and S. Palomas. The very-high-resolution configuration of the ec-earth global model for highresmip. *Geoscientific Model Development*, 18(2): 461–482, 2025. doi: <https://doi.org/10.5194/gmd-18-461-2025>.
- National Aeronautics and Space Administration. Nasa earth observing system data and information system (eosdis). [Online]. URL <https://www.earthdata.nasa.gov/>.
- E. Olason, P. Rampal, and V. Dansereau. On the statistical properties of sea-ice lead fraction and heat fluxes in the arctic. *The Cryosphere*, 15:1053–1064, 2021. doi: <https://doi.org/10.5194/tc-15-1053-2021>.
- R. Pan, Q. Shu, Q. Wang, S. Wang, Z. Song, Y. He, and F. Qiao. Future arctic climate change in cmip6 strikingly intensified by nemo-family climate models. *Geophysical Research Letters*, 50(4), 2023. doi: <https://doi.org/10.1029/2022GL102077>.
- X. Pang, J. Pu, X. Zhao, Q. Ji, M. Qu, and Z. Cheng. Comparison between amsr2 sea ice concentration products and pseudo-ship observations of the arctic and antarctic sea ice edge on cloud-free days. *Remote Sensing*, 10(2):317, 2018. doi: <https://doi.org/10.3390/rs10020317>.

- S. Penny, T. Smith, T. Chen, J. Platt, H. Lin, M. Goodliff, and H. Abarbanel. Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation. *Journal of Advances in Modeling Earth Systems*, 14(3):1–25, 2022. doi: <https://doi.org/10.1029/2021MS002843>.
- M. Peyron, A. Fillion, S. Gurol, V. Marchais, S. Gratton, P. Boudier, and G. Goret. Latent space data assimilation by using deep learning. *Quarterly Journal of the Royal Meteorological Society*, 147(740):3759–3777, 2021. doi: <https://doi.org/10.1002/qj.4153>.
- F. Rabier, H. Jarvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons. The ecmwf operational implementation of four-dimensional variational assimilation. i: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126(564):1143–1170, 2000. doi: <https://doi.org/10.1002/qj.49712656415>.
- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: <https://doi.org/10.1109/CVPR52688.2022.01042>.
- J. Selivanova, D. Iovino, and F. Cocetta. Past and future of the arctic sea ice in high-resolution model intercomparison project (highresmp) climate models. *The Cryosphere*, 18:2739–2763, 2024. doi: <https://doi.org/10.5194/tc-18-2739-2024>.
- Y. Shu, Y. Zhu, F. Xu, L. Gan, P. Lee, J. Yin, and J. Chen. Path planning for ships assisted by the icebreaker in ice-covered waters in the northern sea route based on optimal control. *Ocean Engineering*, 267, 2023. doi: <https://doi.org/10.1016/j.oceaneng.2022.113182>.
- G. Spreen, L. Kaleschke, and G. Heygster. Sea ice remote sensing using amsr-e 89 ghz channels. *Journal of Geophysical Research*, 113:C02S03, 2008. doi: [10.1029/2005JC003384](https://doi.org/10.1029/2005JC003384).
- X. Tian. Jacobian-enforced neural networks (jenn) for improved data assimilation consistency in dynamical models. *arXiv:2412.01013*, 2024. URL <https://arxiv.org/abs/2412.01013>. Accessed: 2025-03-19.
- O. Zavala-Romero, A. Bozec, E. Chassignet, and J. Miranda. Convolutional neural networks for sea surface data assimilation in operational ocean models: test case in the gulf of mexico. *Ocean Science*, 21(1):113–132, 2025. doi: <https://doi.org/10.5194/os-21-113-2025>.
- Q. Zhao, Q. Zhao, S. Peng, J. Wang, S. Li, Z. Hou, and G. Zhong. Applications of deep learning in physical oceanography: a comprehensive review. *Frontiers in Marine Science*, 11, 2024. doi: <https://doi.org/10.3389/fmars.2024.1396322>.

A APPENDIX

A.1 ALGORITHM OF MODIFICATION OF RESTARTS IN THE NEMO OCEAN MODEL

With initial restart files, written by the ocean model and new assimilated fields that came out of assimilation algorithm following was applied:

1. The *sicon* concentration values were clipped to the $[0, 1]$ range to ensure physical consistency. Values below 0.01 were treated as zero, and a mask was created based on this threshold to distinguish ice from open water. The corrected *sicon* data is written to the *a_i* variable.
2. Ice thickness and concentration are used to calculate the ice volume per unit area v_i , using the formula $v_i = a_i \cdot sithic$.
3. Since snow cannot lie on water, the snow volume per unit area (v_s) and the ice concentration (a_i) from the restart file are used to calculate the snow thickness. This snow thickness is then multiplied by the new ice concentration (a_i): $v_s = \frac{v_s[rest] \cdot a_i}{a_i[rest]}$.
4. The *snwice_mass* and *snwice_mass.b* values were recalculated from v_i and v_s using the average densities of sea ice and snow.

5. The salinity variable sv_i is recalculated for the new ice volume v_i . A value of 7.4 psu was assigned for newly formed ice.
6. The values of e_{s_l01} , e_{i_l01} , and e_{i_l02} are recalculated proportionally to v_s and v_i taking into account the heat capacity and latent heat of fusion for snow and ice.
7. The internal stress variables $stress1_i$, $stress2_i$, and $stress12_i$ are recalculated proportionally to the ice volume v_i .
8. The oa_i value is scaled proportionally with a_i .
9. The ocean state components of the restart file were left unmodified, as the change in ice concentration was not substantial.

A.2 ADDITIONAL ASSIMILATION PROCESS ILLUSTRATIONS

These are examples of assimilated sea ice fields for model-to-model assimilation experiments and sat-to-model experiments.

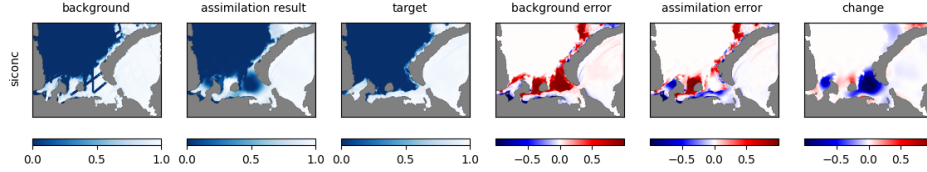


Figure 6: Results M2M assimilation for 12 April with use $3d_var$ model. $sicinc$ — sea ice concentration. Columns: background — initial values, assimilation result — M2M output, target — reference values, background error = background — target (initial field deviation), assimilation error = assimilation result — target (post assimilation deviation), change = assimilation result — background (assimilation induced adjustment).

A.3 THE USE OF LARGE LANGUAGE MODELS (LLMs)

The text in the paper was initially written by humans and then sent to LLM in the process to suggest stylistic improvements and to correct grammar and punctuation mistakes.

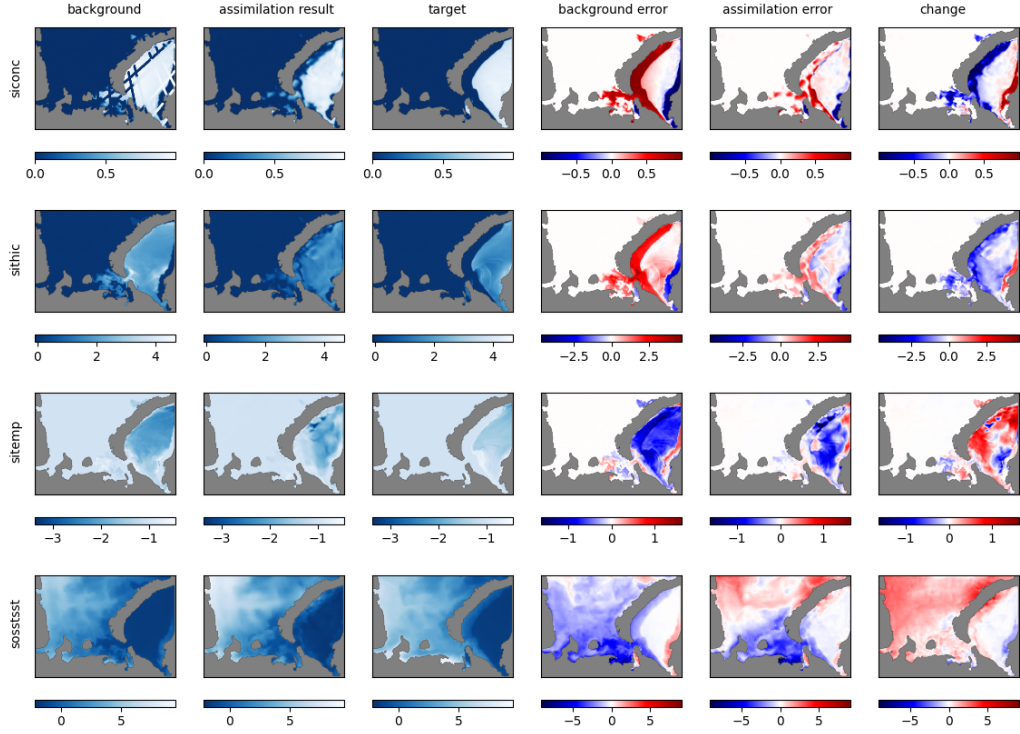


Figure 7: Results M2M assimilation for 30 May with use *vae_4f* model. Rows: *siconc* – sea ice concentration, *sithic* – sea ice thickness, *sosstsst* – sea surface temperature. Columns: background – initial values, assimilation result – M2M output, target – reference values, background error = background – target (initial field deviation), assimilation error = assimilation result – target (post assimilation deviation), change = assimilation result – background (assimilation induced adjustment).

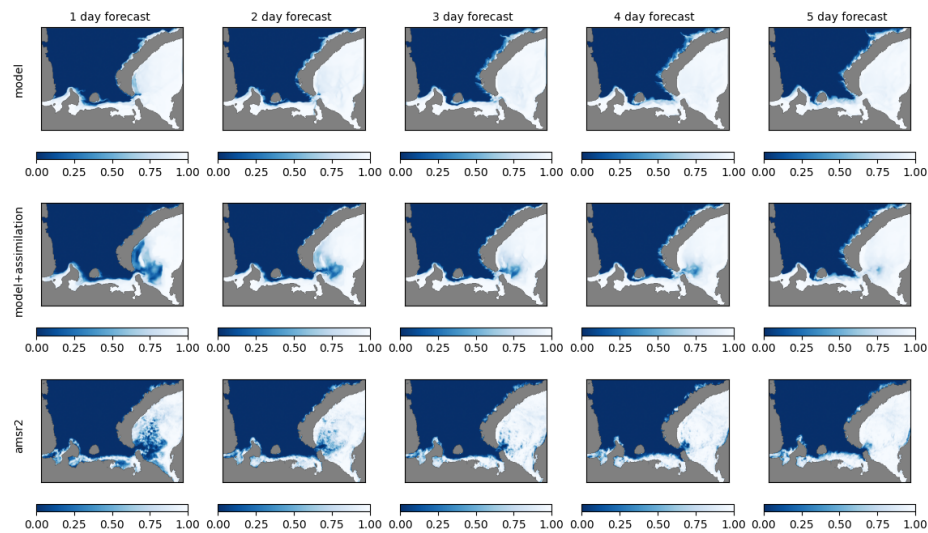


Figure 8: NEMO model prediction results. The forecast was initialized on February 22, 2023, and run for 5 days. Top row: Ice concentration values predicted by the model without data assimilation. Middle row: Ice concentration values predicted by the model with data assimilation (assimilation was performed only on the initial day). Bottom row: Reference AMSR2 satellite ice concentration data for comparison.