DATA-DRIVEN UNCERTAINTY-AWARE FORECASTING OF SEA ICE CONDITIONS IN THE GULF OF OB BASED ON SATELLITE RADAR IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

The increase in Arctic marine activity due to rapid warming and significant sea ice loss necessitates highly reliable, short-term sea ice forecasts to ensure maritime safety and operational efficiency. In this work, we present a novel data-driven approach for sea ice condition forecasting in the Gulf of Ob, leveraging sequences of radar images from Sentinel-1, weather observations, and GLORYS forecasts. Our approach integrates advanced video prediction models, originally developed for vision tasks, with domain-specific data preprocessing and augmentation techniques tailored to the unique challenges of Arctic sea ice dynamics. Central to our methodology is the use of uncertainty quantification to assess the reliability of predictions, ensuring robust decision-making in safety-critical applications. Furthermore, we propose a uncertainty-aware model switching mechanism that enhances forecast accuracy and model robustness, crucial for safe operations in volatile Arctic environments. Our results demonstrate substantial improvements over baseline approaches, underscoring the importance of uncertainty quantification and specialized data handling for effective and reliable sea ice forecasting.

INTRODUCTION

The Arctic region is experiencing an unprecedented rate of warming, leading to a significant reduction in sea ice area by more than 30% over the last four decades, and a simultaneous decrease in sea ice thickness (Kwok, 2018). Alongside this, the last century has seen active development of icebreaker construction technologies, including nuclear-powered ones. These changes have opened up new sea routes, such as the Northern Sea Route, which provide faster and more economical transport. However, increased navigation is accompanied by increased risks due to ice jams, posing a serious threat to maritime safety.

Traditional sea ice models, based on elastic-visco-plastic rheological properties, often fail to accurately reflect all the nuances of ice deformation, rendering them unreliable for forecasting in some cases (Nummelin et al., 2016; Eastwood et al., 2020; Li et al., 2021; Overland & Pease, 1988). Additionally, these models require significant computational resources to adequately simulate the interactions between the ocean and ice. Consequently, there is a need to explore alternative methodologies that leverage statistical methods such as machine learning techniques, known for their flexibility and lower computational demands.

Our research is aimed at improving the forecasting of ice conditions in the Gulf of Ob, a region significantly influenced by the interaction of the saline waters of the Kara Sea and the fresh water of the big northern rivers, leading to complex ice formation dynamics (Weatherly & Walsh, 1996; Osadchiev et al., 2021).

We utilize radar images obtained in the Sentinel-1 mission (Sentinel-1), weather observation data (Weather & Climate), and operational forecasts and reanalysis from the GLORYS project (GLO) to predict future sea ice conditions. From a machine learning perspective, the series of satellite radar images can be treated as a continuous video sequence, therefore the problem can be formulated as a conditioned video prediction task — the widely investigated problem in common-life domain (Ming et al., 2024). Our research employs following video prediction models:

- Implicit Stacked Autoregressive Model for Video Prediction (IAM4VP) (Seo et al., 2023) uses a fully convolutional neural network with an implicit multi-input-single-output workflow, achieving state-of-the-art accuracy of weather predictions in datasets such as SEVIR;
- Dynamic Multi-Scale Voxel Flow Network (DMVFN) (Hu et al., 2023) utilizes voxel flow for video prediction, addressing efficiency and adaptability in handling diverse motion scales;
- MotionRNN (Wu et al., 2021) models spacetime-varying motions using the Motion Gating Recurrent Unit and Motion Highway mechanisms, enhancing prediction accuracy in dynamic scenarios;
- Neural Ordinary Differential Equations (Neural ODE) and Vid-ODE (Park et al., 2021) treat consecutive frames as solutions to systems of ordinary differential equations, offering control over visual attributes and smooth transitions between frames.

As the primary loss for training models and metric for evaluating their performance, we use Mean Squared Error (MSE) between predicted and target Synthetic-Aperture Radar (SAR) images. In addition to this, we utilize the Structural Similarity Index (SSIM) (Wang et al., 2004) and its extension, the Multi-Scale Structural Similarity Index (MS-SSIM) (Wang et al., 2003), to assess the perceived quality of digital images and videos. Finally, the Integrated Ice Edge Error at level c (IIEE@c) (Goessling et al., 2016) is utilized to measure the similarity between forecasted and observed ice sheets. These indicators allow us to meticulously compare the accuracy of predicted ice conditions against observed data, ensuring that our models reflect not only general trends but also detailed spatial structures necessary for accurate ice mapping.

Our contributions can be summarized as follows:

- we explore the potential of modern deep learning video-prediction models in short-term regional sea ice forecasting;
- we address the problem of data irregularity and missing values within the Arctic area by exploring filtration, normalization, and augmentations techniques;

- we show the ensemble of ML models provides sufficient uncertainty estimation, and we propose novel uncertainty-aware model switching scheme that stabilizes the forecast and enhances its quality;
- finally, we assess a gap filling performance for satellite radar imagery and demonstrate the superiority of our method in comparison with a general approach for interpolating video sequences.

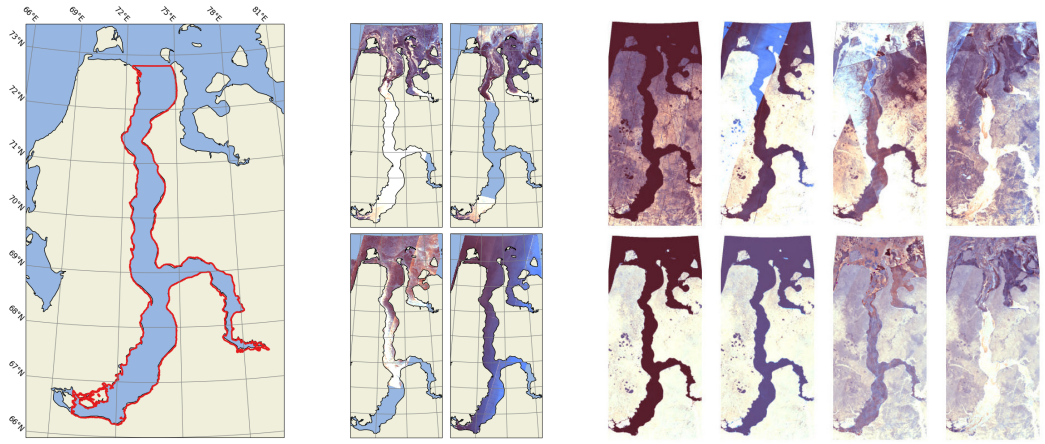
RELATED WORKS

Several studies have demonstrated the effectiveness of machine learning in forecasting sea ice extent and sea ice concentration. Chi and Kim (Chi & Kim, 2017) pioneered in the use of deep learning for sea ice prediction. Their model employs multilayer perceptrons (MLPs) and long- and short-term memory networks (LSTMs) to capture complex relationships in sea ice data. By training the MLP- and LSTM-based models on historical data, they identify patterns for one-month predictions, outperforming traditional statistical models. This work highlights the advantages of deep learning in sea ice forecasting.

Recent research has extended the application of UNet-based models to sea ice forecasting, highlighting their versatility beyond original medical imaging applications. Fernandez et al. (Fernández et al., 2022) investigated coastal sea elements forecasting using various UNet-based architectures, including 3DDR-UNet and its enhanced versions. Their study demonstrated the effectiveness of these models in forecasting coastal sea conditions when using satellite imagery. Grigoryev et al. (Grigoryev et al., 2022) presented a recurrent UNet with a specialized training scheme that considerably outperformed persistence and linear trend baseline forecasts of sea-ice conditions in the regions of the Barents, Labrador, and Laptev seas for lead times up to 10 days. Kvanum et al. (Kvanum et al., 2024) showed that the similar approach in the Barents sea can overcome traditional numerical models at the forecasting of sea ice concentration at one kilometer resolution and 3-day lead time. Keller et al. (Keller et al., 2023) explored various UNet-based architectures for prediction sea ice extent at kilometer resolution for lead time up to 7 days. These studies revealed the potential of machine learning methods over traditional approaches for high-resolution sea ice conditions forecasting.

Several studies showcase the prospects of uncertainty-aware data-driven sea ice forecasting. Horvath et al. (Horvath et al., 2020) suggested using Bayesian logistic regression to forecast September minimum ice cover from 1-month up to 7-month lead times. In this paper Bayesian uncertainty quantification helps to assess the reliabillity of the forecasts. Andersson et al. (Andersson et al., 2021) introduced a probabilistic deep learning sea ice forecasting system called IceNet with 6-month lead time. Their system predicts monthly averaged sea ice concentration maps at 25-km resolution, outperforming traditional models by effectively bounding the ice edge. Wu et al. (Wu et al., 2022) suggested VAE-Based Non-Autoregressive Transformer as an uncertainty-aware model for long-term sea ice concentration forecast along Northern Sea Route. Also uncertainty quantification is the key motive of the conjugate problems like sea ice data assimilation (Nazanin, 2019) or sea ice concentration retrieval (Chen et al., 2023b).

One challenge in sea ice forecasting and analysis based on satellite imagery data is the presence of noise and gaps, which can occur due to limitations of satellite trajectories, instrumental errors, data losses, and environmental factors. To address this, researchers have proposed various gap-filling methods (Desai & Ganatra, 2012). They incorporate chained data fusion, multivariate interpolation, and empirical orthogonal functions to effectively fill missing data. Weiss et al. (Weiss et al., 2014) proposed an effective approach for continental-scale gaps inpainting based on nearest neighbors method and taking in account seasonality of the data. Their approach, additionally, quantified uncertainty of the filled values. Appel (Appel, 2024) introduced a deep learning approach based on partial convolutions for filling gaps in consecutive data, highlighting the promise of deep learning for satellite imagery-related tasks. These methods enhance the quality and reliability of remote observations, having potentially a wider range of applications than just sea ice analysis.



(a) (b) (c)

Figure 1: (a) Gulf of Ob region we are focused on. (b) Examples of colorized SAR images. (c) Examples of images before filtration (the first line) and after (the second line). Images are colorized according to sentinelhib guidelines.

DATA

Our neural network model utilizes a number of input channels (fields) that come from three sources: Sentinel-1 (Sentinel-1) SAR imagery in HV and HH polarizations in extra-wide mode, Global Ocean Physics Reanalysis (GLORYS) (GLO), and historical data from meteostations (Weather & Climate). See detailed information in the appendix A. For the purpose of this study we set the target resolution to one kilometer, which nevertheless is sufficient enough for navigation applications (Kvanum et al., 2024; Keller et al., 2023).

The region we investigated encompasses the Gulf of Ob and the Taz Estuary in Northern Russia. The region of interest, at this one kilometer resolution, produces images with a size of 880×400 pixels. SAR images are interpolated conservatively to a covering equal-area grid in North-Polar projection (see Figure 1). GLORYS is interpolated bilinearly. Meteodata is interpolated using radial basis function interpolation. To focus on forecasting sea-ice dynamics, the land surface in target images is masked with zero values.

In this study, we selected SAR imagery as the target, due to several key advantages it offers. Firstly, Sentinel-1 allows for continuous monitoring of polar regions regardless of cloud cover or illumination. Secondly, the high spatial resolution of Sentinel-1 enables detailed analysis of sea ice, including the detection of small-scale ice features important for navigation and environmental monitoring. Thirdly, the large amount of historical data provided by Sentinel-1 is essential for training deep learning models in comparison with others sources. Finally, the almost real-time data delivery of Sentinel-1 is crucial for operational applications.

We acknowledge several disadvantages of the SAR imagery. Firstly, the revisiting period of the satellite is several days, hence many empty frames to appear when attempting to create a regular time

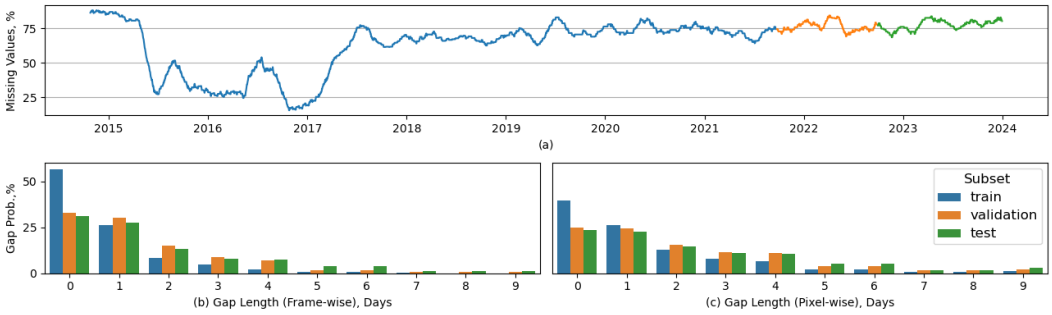


Figure 2: (a) Frequency of missing values in SAR imagery smoothed with a month-wide rolling window. (b-c) Distribution of distances between consecutive missing values across all subsets, calculated frame-wise between frames with any valid data, and pixel-wise at fixed locations.

series sequence. The distribution of missing values over subsets is depicted in Figures 2. Secondly, the entire area is not always captured in the images, resulting in some images being incomplete. Thirdly, thermal noise and imagery artifacts at HV polarization are significant, leading to varying brightness in repetitive patterns known as scalloping.

METHODS

DATA PREPROCESSING AND FILTRATION

The origin of the noise in SAR imagery is thermal interference within radar systems, influenced by the technology utilized for surface scanning, resulting in presence of speckles and scalloping patterns (Singh et al., 2021). Thermal artifacts have significant magnitude relative to useful information, which leads to huge biases and corrupts optimization convergence of neural networks. Therefore, we preprocess data to filter out imagery artifacts. The results of the final filtering are presented in Figure 1c.

Our custom filtering algorithm treats images as vectors from $\mathbb{R}^{H\times W}$ space with standard scalar product, where H and W stand for sizes of the input frames. The core assumption is the orthogonality of artifacts A to the subspace of clear images $C \perp A$. Therefore, the filtering process is an orthogonal projection: $P : \mathbb{R}^{H\times W} \to C$, $P^2 = P = P^T$.

However, the construction of such operator requires the retrieval of aforementioned subspaces. The key thought is that all the ice-free frames IF must have the same projection: $\exists c_0 \, \forall c \in IF : P(c) = c_0$. To achieve this, the frame c_0 with neither ice nor noise should be chosen by hand. We obtained several candidates for such a frame through visual assessment and peaked the pixel-wise minimum of all of them. Then artifact subspace A is constructed from IF to match orthogonality condition to c_0 at least, and the filtering operator P is constructed after choosing a basis in the subspace containing all the artifacts A:

$$A = \{v - \frac{(v, c_0)}{(c_0, c_0)} c_0 | v \in IF\}, \quad P(v) = v - \sum_{i=1}^{n} \frac{(v, e_i)}{(e_i, e_i)} e_i \tag{1}$$

where $\{e_i\}_{i=1}^n$ is a basis in the linear span of A.

VIDEO PREDICTION MODELS

To determine the relative quality of our models performance, we compare them against two baselines: persistence forecast and linear one. To obtain the parameters of the linear transformation we utilize the same techniques as for deep learning models.

IAM4VP (Seo et al., 2023) is fully convolutional neural network that leverages the trade-off between temporal-consistency of autoregressive methods and error-independence of non-autoregressive ones via implicit Multi-Input-Single-Output workflow. Like non-autoregressive methods, stacked autoregressive approach uses the same observed sequence to estimate future frames. However, the model uses its own predictions as input, similar to autoregressive methods. As the number of time steps increases, predictions are sequentially stacked in the queue. After the iterative process is finished, all generated frames are refined by the last few layers to raise the temporal correlations.

DMVFN (Hu et al., 2023) is a video prediction model leveraging voxel flow estimation to focus on movement and to handle the occlusion effect. DMVFN also incorporates a dynamic routing module that adaptively selects sub-networks based on the input frames, allowing it to handle diverse motion scales efficiently. The model's architecture includes Multi-scale Voxel Flow Blocks (MVFBs) that capture large motions and iteratively refine voxel flow estimates. DMVFN demonstrates improved efficiency and adaptability, particularly for videos with complex motion patterns and is considered a state-of-the-art deep learning solution for video prediction.

MotionRNN (Wu et al., 2021) is a model designed for video prediction, specifically addressing the challenge of predicting continuous spatio-temporal dynamics. MotionRNN is a successor of the LSTM-based architectures that also incorporates warp transformation and introduces the concept of breaking down physical motions into transient variation and motion trend. Transient variation

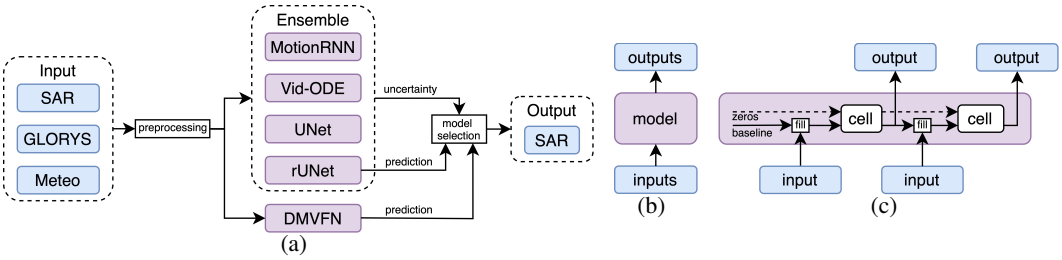


Figure 3: (a) The overall pipilene: data gathering, interpolation, normalization, filtration, neural networks evaluation, uncertainty quantification and the model selection. The final prediction is chosen from outputs of rUNet and DMVFN, based on the uncertainty, estimated by ensemble spread. (b) Non-autoregressive models treat time data as image channels and predicts output at fixed number of lead times. (c) Autoregressive models make forecast day by day, the intermediate forecast is used for missing values imputation, while the missing values at the first timestamp are imputed with persistent baseline.

represents immediate changes, while the motion trend captures the overall direction or tendency of movements over time.

Neural ODE (Chen et al., 2018) offers a powerful framework for modeling dynamic systems by means of machine learning. The core idea is to model the transformation of data through a continuous dynamic equation in a Cauchy formulation instead of discrete layers used in traditional neural networks. The forward pass is a numerical solution of a parametrized ODE. To train Neural ODEs through backpropagation the gradients can be computed either directly through the dynamic equation using automatic differentiation or more memory-efficiently through the adjoint method. The adjoint method treats gradients as solutions of a reverse-time differential equations, integrating it backwards in time.

Vid-ODE (Park et al., 2021) represents Neural ODE in a latent space on motion-vector dynamics with a warp-correction mechanism. The main idea of Vid-ODE lies in the parameterizing visual attributes such as pixels position by utilizing differential equations. The encoded input frames are assimilated into an evolving state by the GRU cell. Neural ODE models latent dynamics. The predicted state is decoded into pixels relative shifts and correction features that are used to correct warping impurities and model color and brightness variation. This iterative process ensures smooth transitions between generated frames.

UNet (Ronneberger et al., 2015), is an convolutional neural network which includes an encoder for capturing context and a decoder for precise localization for the output. The decoder path involves upsampling of the feature maps and concatenates them with the corresponding feature maps from the encoder path. Originally designed for the biomedical image segmentation, UNet has been adapted for different geophysical fields forecasts such as: coastal sea elements (Fernández et al., 2022), precipitation (Kaparakis & Mehrkanoon, 2023; Trebing et al., 2021), and sea ice concentration (Grigoryev et al., 2022; Kvanum et al., 2024). We use it within both autoregressive and non-autoregressive approaches. The former one we will mention as rUNet, where 'r' stands for recurrent.

Data split

The data is divided into three sets. Training set: September 1, 2015, to September 23, 2021. Validation set: September 24, 2021, to September 30, 2022; Test set: October 1, 2022, to September 30, 2023.

Implementation and training

All models are implemented in PyTorch and trained from scratch with AdamW optimizer (Paszke et al., 2019). The loss function is a combination of MSE and SSIM losses:

$$L = \text{MSE} - 0.2 \cdot \text{SSIM} \tag{2}$$

Table 1: Model configurations. Regime abbreviations are constructed as follows: SI stands for Single-Input, MI — Multi-Input, SO — Single-Output, and MO — Multi-Output, based on sequence lengths: Single-Input models acquire input iteratively, Multi-Input — at once; Single-Output models are autoregressive predictors; Multi-Output — non-autoregressive. Computational costs for each model (in GFLOPS) are provided per one input sequence. ODE-based models use adaptive solvers; the adaptive time step leads to varying GFLOPS; its standard deviation is provided.

Model	Regime	Input size	GFLOPS	Params
Persistence	-	-	-	0
Linear	SISO	7	33	1.06 K
DMVFN	MISO	7	198	3.56 M
IAM4VP	Implicit MISO	10	76	27.8 M
Neural ODE	SISO	7	200 ± 100	18.53 K
MotionRNN	SISO	7	10610	6.84 M
Vid-ODE	SISO	7	480 ± 150	469 K
UNet	MIMO	7	559	31.10 M
rUNet	SISO	7	4780	31.04 M

The initial learning rate is set to 10^{-3} and exponentially decreasing with factor $\gamma = 0.99$. The batch size is set to 32. Models are trained until either convergence of validation metrics or the overfitting begins (early-stopping). Models with the best validation score are evaluated after on the test-subset.

To mitigate bias on missed parts of the SAR input, normalization layers were removed from encoders of IAM4VP and UNet. For other models missing values are imputed with previous forecast from autoregressive prediction (see the schematic representation on Figure 3).

While training Neural ODE and Vid-ODE models, naive implementations of the adjoint method might suffer from inaccuracy in reverse-time trajectory computation, therefore in our work we have used specific implementation called MALI (Zhuang et al., 2021) that guarantees accuracy in gradient estimation.

To determine the relative quality of our models performance, we compare them against two baselines: persistence forecast and linear one. To obtain the parameters of the linear transformation we utilize the same techniques as for deep learning models. The overall models configurations are provided in Table 1.

AUGMENTATIONS

To prevent overfitting and improve generalization ability we utilize geometrical augmentations: random horizontal flips with a probability of 0.5 and uniformly sampled random rotations with angles in range $[-10°, 10°]$ with the corresponding rotation of wind and sea-currents field. To leverage the imbalance of missing values depicted at Figure 2 we utilized frameout augmentation. Up to three random frames in the input sequence are cut out until the concentration of missing values reaches the level of the test subset (70%).

UNCERTAINTY-AWARENESS

Estimating uncertainty in data-driven weather forecasting models is crucial for better model interpretation and decision-making. If the uncertainty estimation is well-calibrated, the reliable predictions are characterized by high confidence. On the other hand, low confidence means the prediction can not be trusted. In such cases one could replace it with a simple baseline or a more robust model. Following this principle, automatized pipelines of uncertainty-aware model mixture can be designed (Lakshminarayanan et al., 2017; Chen et al., 2023a; Zeng et al., 2023; Jiang et al., 2023). The mechanism is as follows: the expert model makes a prediction, its uncertainty is estimated; if the uncertainty exceeds the preset threshold, the prediction is replaced by more stable baseline. In our work the threshold is selected on the validation subset. This helps to exclude unreliable forecasts and enhances the overall performance of the forecasting system.

Traditional weather and climate models estimate uncertainty as the spread of an ensemble, constructed by the model inputs perturbations (Grimit & Mass, 2007). The ensemble spread is defined

as a standard deviation of predictions. Previous research (Scher & Messori, 2021) showed that, when using neural networks, ensembles of models with similar architectures (homogeneous) provide similar results. Models weights in the ensemble have to be perturbed with retraining, dropout, etc. Moreover, there are premises that an ensemble of diverse architectures (heterogeneous) might provide better uncertainty estimation (Zaidi et al., 2022).

In our research we construct both homogeneous and heterogeneous ensembles and compare their spread as a predictor for the uncertainty estimation for the model selection mechanism. The suggested pipeline does not impose additional costs as all the models do not need to be modified or retrained.

RESULTS

FORECASTING

While designing the experiments, we focused on evaluating the performance and stability of various forecasting models. Our results reveal a trade-off between achieving high computer vision metrics and maintaining forecast stability — while some models excel in certain metrics, their forecasts can be less consistent. However, we found that an ensemble of four high-performing models with diverse architectures — namely MotionRNN, Vid-ODE, UNet, and rUNet — offers robust uncertainty estimation. The most significant improvement over the baseline across nearly all metrics was achieved using a uncertainty-aware model switching scheme which utilized an rUNet backbone, an autoregressive UNet, and DMVFN as a robust alternative (see Figure 3a for schematic representation).

A summary of the metrics evaluated on the test subset for all trained models with uncertainty-aware model switching is presented in Table 2. Figure 4 shows detailed improvements over the baseline, broken down by month and lead time. Figure 13 illustrates the detailed RMSE by date for individual models, along with the corresponding ensemble spreads for several configurations. Examples of predictions are provided in Figure 11. Using the mean of ensembles instead of model selection yields only a marginal improvement in the final metrics, as shown in Table 3.

Table 2: Summary of the test metrics (lower is better) for models with confidence-based mixture with DMVFN as a robust model; the uncertainty is estimated by the ensemble spread of predictions from MotionRNN, Vid-ODE, UNet, and rUNet models. The standard deviation for the best model (rUNet) is estimated by training with three random initializations.

Model	**MSE**	**1 - SSIM**	**1 - MS-SSIM**	**IIEE@15**	**IIEE@30**	**IIEE@50**	**IIEE@75**
		$(\times 10^{-3})$			$(\times 10^{-2})$		
Persistence	11.2	9.8	5.6	11.5	10.4	11.0	7.3
Linear	9.8	9.1	5.2	14.0	9.6	11.1	7.7
DMVFN	10.0	8.8	5.1	11.7	10.2	10.8	6.9
IAM4VP	8.7	10.5	5.5	14.7	10.6	11.0	7.1
Neural ODE	8.3	9.3	4.9	12.1	10.1	10.7	6.2
MotionRNN	7.3	9.0	4.7	11.4	9.3	9.9	5.9
Vid-ODE	7.5	8.7	4.7	12.1	9.2	9.7	5.7
UNet	7.7	**8.2**	4.6	12.1	9.3	9.6	6.0
rUNet	**6.8**$\pm$0.2	8.3$\pm$0.2	**4.5**$\pm$0.1	**10.0**$\pm$1.1	**9.0**$\pm$0.3	**9.2**$\pm$0.2	**5.3**$\pm$0.2

Table 3: Summary of the test metrics for ensembles. uncertainty-aware model switching to DMVFN is utilized. "rUNet x3" stands for mean forecast of three retrained versions of rUNet. "Best 4" stands for mean prediction from MotionRNN, Vid-ODE, UNet, and rUNet models.

Ensemble	**MSE**	**1 - SSIM**	**1 - MS-SSIM**	**IIEE@15**	**IIEE@30**	**IIEE@50**	**IIEE@75**
		$(\times 10^{-3})$			$(\times 10^{-2})$		
rUNet x3	6.7	8.3	4.5	10.0	8.9	9.1	5.3
Best 4	6.6	8.2	4.4	11.2	8.7	9.3	5.2

Following the Grigoryev's work (Grigoryev et al., 2022), models trained to produce 3-day forecasts were also tested with 10-day outputs without any additional fine-tuning. The results are presented

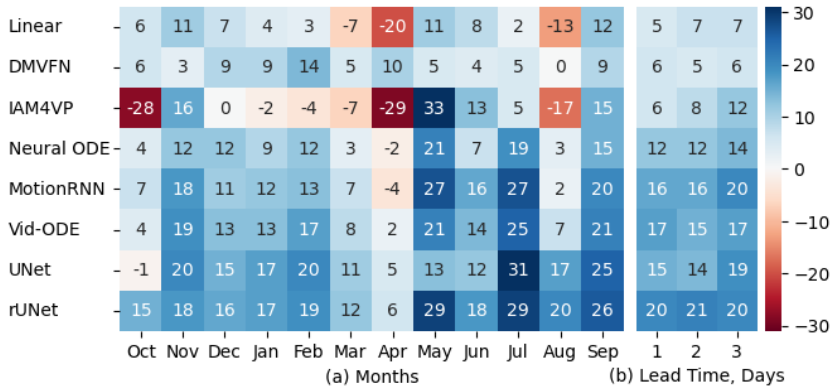


Figure 4: The RMSE percentage improvement over persistence baseline for each month (a), and for each lead time in days (b), over the test subset. The colormap is shared.

in Figure 9. Linear and ODE-based models accumulate errors exponentially, degrading over persistence after the 5-th day. Other models errors increase linearly over time, providing stable improvement over persistence, except IAM4VP which manages to overcome all other models after the 7-th day of the forecast.

Due to the irregular intervals at which satellites capture the target region, a strong correlation between model performance and the length of gaps between valid images is expected. This relationship is illustrated in Figure 10, where RMSE generally increases linearly until the gap length surpasses the models' input size. Once this threshold (7 days) is exceeded, the RMSE approaches that of the persistence baseline.

GAP FILLING

The developed pipeline is particularly useful for filling gaps in SAR images, a common issue in satellite data. Building on the approach proposed by Appel (Appel, 2024), this gap-filling process can be effectively performed as a 1-day forecast. To improve accuracy, we combine forward and backward forecasts, denoted as y_F and y_B, respectively. By incorporating the uncertainty estimates of these forecasts, σ_F and σ_B, we can weight them appropriately and calculate an overall uncertainty using a harmonic mean:

$$y = \frac{\sigma_B}{\sigma_F + \sigma_B} y_F + \frac{\sigma_F}{\sigma_F + \sigma_B} y_B, \quad \sigma = \frac{2\sigma_F \sigma_B}{\sigma_F + \sigma_B} \tag{3}$$

A key advantage of this approach is that it does not require retraining the models. We evaluated the performance of this gap-filling method using a leave-one-out cross-validation technique (Kohavi, 1995). For comparison, we also tested the pretrained AdaCoF model (Lee et al., 2020), which is one of the state-of-the-art models for video interpolation. As shown in Table 4, our pipeline achieved a strong R^2 value of 87.7%. This is consistent with similar R^2 values reported in the literature (Weiss et al., 2014; Appel, 2024) for gap-filling in satellite imagery under similar conditions, such as missing swaths up to 500 km wide and 1 kilometer resolution.

Table 4: Summary of gap filling metrics obtained during a leave-one-out validation. Uncertainty-aware mixture of rUNet and DMVFN is utilized for forward and backward forecasts, where the "Forward+Backward" is a uncertainty-weighted mean. The input channels related to wind and currents are reversed for the backward run. The best metric values in each column are highlighted in bold.

Model	**MSE**	**1 - SSIM**	**1 - MS-SSIM**	**IIEE@15**	**IIEE@30**	**IIEE@50**	**IIEE@75**
	($\times 10^{-3}$)			($\times 10^{-2}$)			
AdaCoF	7.3	8.3	4.5	**8.0**	7.9	9.2	6.0
Forward	6.5	8.1	4.4	8.6	8.5	9.2	**5.3**
Forward+Backward	**6.0**	**7.6**	**4.1**	9.1	**7.7**	**8.7**	5.7

DISCUSSION

This research addresses the critical challenge of short-term regional sea ice forecasting, exploring a variety of approaches to improve prediction accuracy and reliability. Among the methods investigated, modern deep learning models for video prediction were tested for their potential in forecasting sea ice dynamics. However, the performance of these models is constrained by several factors, including the scarcity of high-resolution data, the complex physical processes governing sea ice behavior, the stochastic nature of daily ice dynamics, and the discontinuities present in ice sheet structures. We argue that these domain specialties mostly affect motion-related elements of video prediction models like flow estimation and prediction, see appendix C for further details.

UNet-based models deliver the best individual results, whereas state-of-the-art video prediction models struggle to surpass baseline performance, though they do offer varying levels of stability. It could be argued that the DMVFN model fails to accurately reproduce sea ice thermodynamics due to its architectural limitations, which, paradoxically, contribute to more stable forecasts. On the other hand, IAM4VP, while efficient at modeling different dynamics with minimal computational cost, produces the most unstable predictions, likely due to the lack of sufficient training data.

Advanced use of uncertainty-aware model switching scheme can further enhance the metrics. The ensemble spread of heterogeneous architectures provides accurate uncertainty estimation for the forecasted fields. Although the model-selection mechanism reduces the final spread-error correlation, the total variance in model error can still be explained up to 87% by accounting for the sea ice concentration and its rate of change (see Figure 14).

CONCLUSIONS

In this research article, we address the challenge of predicting ice conditions in the Gulf of Ob, a region characterized by complex ice formation dynamics influenced by the interaction of saline water and freshwater. We explore the potential of machine learning methods as an alternative to traditional numerical sea ice models, aiming to improve forecasting accuracy and efficiency.

Our key findings reveal that even modern state-of-the-art machine learning models can not achieve sufficient forecasting performance solely. Furthermore, domain-aware data preprocessing and augmentations are essential to train deep learning models for this task. All models struggle due to lack of training data, long gaps in it and complex sea ice dynamics, leading to tricky fidelity-stability trade-off. Although usage of ensembles cannot significantly improve average models performance, it helps to eliminate high errors due to outliers in data, especially in spring season, thus increasing overall system reliability. We consider also an interesting finding that the different ML models capture different aspects of the ice dynamics in such a way that their ensemble gives a reliable forecast uncertainty quantification, as the spread-error correlation coefficient reaches 87%. To overcome the aforementioned trade-off we construct the uncertainty-aware model switching scheme, that provides both stable and explainable forecasts while improving general performance. The mixture of the rUNet and DMVFN architectures provides the best computer vision and geophysical metrics and beats baselines by a wide margin.

Future research directions include developing models that can effectively capture the dynamics of ice formation and melting is crucial. Additionally, addressing the limitations of current approaches through more advanced architectures and techniques can also be beneficial. Further advancements in sea ice forecasting will not only improve maritime navigation safety but also deepen our understanding of complex sea ice dynamics.

REPRODUCIBILITY

The developed source code was attached to the manuscript as a supplementary material. The preprocessed dataset is available upon written request. The processing procedure and the training of the models is described in the Methods section of the paper. Both code and dataset will be published and available after the end of double-blind review.

REFERENCES

European union-copernicus marine service. glorys12v1. URL https://resources.marine.copernicus.eu/product-detail/GLOBAL_ANALYSISFORECAST_PHY_001_024/INFORMATION.

Tom R Andersson, J Scott Hosking, María Pérez-Ortiz, Brooks Paige, Andrew Elliott, Chris Russell, Stephen Law, Daniel C Jones, Jeremy Wilkinson, Tony Phillips, et al. Seasonal arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(1):5124, Aug 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25257-4.

Marius Appel. Efficient data-driven gap filling of satellite image time series using deep neural networks with partial convolutions. *Artificial Intelligence for the Earth Systems*, 3(2), apr 2024. ISSN 2769-7525. doi: 10.1175/aies-d-22-0055.1. URL http://dx.doi.org/10.1175/AIES-D-22-0055.1.

Annie Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. Confidence-based model selection: When to take shortcuts for subpopulation shifts, 06 2023a.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Xinwei Chen, Ray Valencia, Armina Soleymani, and K. Andrea Scott. Predicting sea ice concentration with uncertainty quantification using passive microwave and reanalysis data: A case study in baffin bay. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023b. doi: 10.1109/TGRS.2023.3250164.

Junhwa Chi and Hyun-choel Kim. Prediction of arctic sea ice concentration using a fully data driven deep neural network. *Remote Sensing*, 9(12):1305, dec 2017. ISSN 2072-4292. doi: 10.3390/rs9121305. URL http://dx.doi.org/10.3390/rs9121305.

Manali Desai and Amit Ganatra. Survey on gap filling in satellite images and inpainting algorithm. *International Journal of Computer Theory and Engineering*, pp. 341–345, 2012. ISSN 1793-8201. doi: 10.7763/ijcte.2012.v4.479. URL http://dx.doi.org/10.7763/IJCTE.2012.V4.479.

Rosemary Eastwood, R. Macdonald, Jens Ehn, Joel Heath, L. Arragutainaq, Paul Myers, D. Barber, and Zou Zou Kuzyk. Role of river runoff and sea ice brine rejection in controlling stratification throughout winter in southeast hudson bay. *Estuaries and Coasts*, 43, 03 2020. doi: 10.1007/s12237-020-00698-0.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson (eds.), *Image Analysis*, pp. 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45103-7.

Jesús García Fernández, Ismail Alaoui Abdellaoui, and Siamak Mehrkanoon. Deep coastal sea elements forecasting using unet-based models. *Knowledge-Based Systems*, 252:109445, sep 2022. ISSN 0950-7051. doi: 10.1016/j.knosys.2022.109445. URL http://dx.doi.org/10.1016/j.knosys.2022.109445.

H. F. Goessling, S. Tietsche, J. J. Day, E. Hawkins, and T. Jung. Predictability of the arctic sea ice edge. *Geophysical Research Letters*, 43(4):1642–1650, 2016. ISSN 1944-8007. doi: 10.1002/2015gl067232. URL http://dx.doi.org/10.1002/2015GL067232.

Timofey Grigoryev, Polina Verezemskaya, Mikhail Krinitskiy, Nikita Anikin, Alexander Gavrikov, Ilya Trofimov, Nikita Balabin, Aleksei Shpilman, Andrei Eremchenko, Sergey Gulev, Evgeny Burnaev, and Vladimir Vanovskiy. Data-driven short-term daily operational sea ice regional forecasting. *Remote Sensing*, 14(22), 2022. ISSN 2072-4292. doi: 10.3390/rs14225837. URL https://www.mdpi.com/2072-4292/14/22/5837.

Eric P. Grimit and Clifford F. Mass. Measuring the ensemble spread–error relationship with a probabilistic approach: Stochastic ensemble results. *Monthly Weather Review*, 135(1):203 – 221, 2007. doi: 10.1175/MWR3262.1. URL https://journals.ametsoc.org/view/journals/mwre/135/1/mwr3262.1.xml.

Sean Horvath, Julienne Stroeve, Balaji Rajagopalan, and William Kleiber. A bayesian logistic regression for probabilistic forecasts of the minimum september arctic sea ice cover. *Earth and Space Science*, 7(10), 2020. doi: 10.1029/2020EA001176.

Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6131, 2023. doi: 10.1109/CVPR52729.2023.00593.

Yuchang Jiang, Vivien Sainte Fare Garnot, Konrad Schindler, and Jan Dirk Wegner. Mixture of experts with uncertainty voting for imbalanced deep regression problems, 2023.

Christos Kaparakis and Siamak Mehrkanoon. Wf-unet: Weather fusion unet for precipitation nowcasting, 2023. URL https://arxiv.org/abs/2302.04102.

Mary Ruth Keller, Christine Piatko, Mary Versa Clemens-Sewall, Rebecca Eager, Kevin Foster, Christopher Gifford, Derek Rollend, and Jennifer Sleeman. Short-term (7 day) beaufort sea ice extent forecasting with deep learning. *Artificial Intelligence for the Earth Systems*, 2(4):e220070, 2023. doi: 10.1175/AIES-D-22-0070.1.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pp. 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603638.

A. F. Kvanum, C. Palerme, M. Müller, J. Rabault, and N. Hughes. Developing a deep learning forecasting system for short-term and high-resolution prediction of sea ice concentration. *EGUsphere*, pp. 1–26, 2024. doi: 10.5194/egusphere-2023-3107. URL https://egusphere.copernicus.org/preprints/2024/egusphere-2023-3107/.

R Kwok. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environmental Research Letters*, 13(10):105005, 2018. ISSN 1748-9326. doi: 10.1088/1748-9326/aae3ec. URL http://dx.doi.org/10.1088/1748-9326/aae3ec.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

S. A. Lapin, E. L. Mazo, and P. N. Makkaveev. Integrated research on the gulf of ob (july to october 2010). *Oceanology*, 51(4):711–715, aug 2011. ISSN 1531-8508. doi: 10.1134/s0001437011040096. URL http://dx.doi.org/10.1134/S0001437011040096.

Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, jun 2020. doi: 10.1109/cvpr42600.2020.00536. URL http://dx.doi.org/10.1109/CVPR42600.2020.00536.

Ming Li, Ren Zhang, and Kefeng Liu. Machine learning incorporated with causal analysis for short-term prediction of sea ice. *Frontiers in Marine Science*, 8, may 2021. ISSN 2296-7745. doi: 10.3389/fmars.2021.649378. URL http://dx.doi.org/10.3389/fmars.2021.649378.

Valentin Ludwig, Gunnar Spreen, and Leif Toudal Pedersen. Evaluation of a new merged sea-ice concentration dataset at 1 km resolution from thermal infrared and passive microwave satellite data in the arctic. *Remote Sensing*, 12(19), 2020. ISSN 2072-4292. doi: 10.3390/rs12193183. URL https://www.mdpi.com/2072-4292/12/19/3183.

Ruibo Ming, Zhewei Huang, Zhuoxuan Ju, Jianming Hu, Lihui Peng, and Shuchang Zhou. A survey on video prediction: From deterministic to generative approaches, 2024.

Asadi Nazanin. *Data-driven Regularization and Uncertainty Estimation to Improve Sea Ice Data Assimilation*. PhD thesis, University of Waterloo, 2019. URL http://hdl.handle.net/10012/14770.

Kimura Noriaki, Nishimura Akira, Tanaka Yohei, and Yamaguchi Hajime. Influence of winter sea-ice motion on summer ice cover in the arctic. *Polar Research*, 32, 2013. doi: 10.3402/polar.v32i0.20193. URL https://polarresearch.net/index.php/polar/article/view/3087.

Aleksi Nummelin, Mehmet Ilicak, Camille Li, and Lars H. Smedsrud. Consequences of future increased arctic runoff on arctic ocean stratification, circulation, and sea ice cover. *Journal of Geophysical Research: Oceans*, 121(1):617–637, 2016. doi: 10.1002/2015JC011156. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC011156.

Alexander Osadchiev, Olga Konovalova, and Alexandra Gordey. Water exchange between the gulf of ob and the kara sea during ice-free seasons: The roles of river discharge and wind forcing. *Frontiers in Marine Science*, 8, dec 2021. ISSN 2296-7745. doi: 10.3389/fmars.2021.741143. URL http://dx.doi.org/10.3389/fmars.2021.741143.

James E. Overland and Carol H. Pease. Modeling ice dynamics of coastal seas. *Journal of Geophysical Research: Oceans*, 93(C12):15619–15637, 1988. doi: 10.1029/JC093iC12p15619.

Sunghyun Park, Kangyeol Kim, Junsoo Lee, Jaegul Choo, Joonseok Lee, Sookyung Kim, and Edward Choi. Vid-ode: Continuous-time video generation with neural ordinary differential equation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2412–2422, may 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i3.16342. URL http://dx.doi.org/10.1609/aaai.v35i3.16342.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pp. 234–241. Springer International Publishing, 2015. ISBN 9783319245744. doi: 10.1007/978-3-319-24574-4_28. URL http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Sebastian Scher and Gabriele Messori. Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, 13(2), 2021. doi: 10.1029/2020MS002331.

Sentinel-1. Copernicus sentinel data 2024, processed by esa. URL https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar.

Sentinel Hub. URL https://custom-scripts.sentinel-hub.com/sentinel-1/sar-ice/.

Min-seok Seo, Hakjin Lee, Doyi Kim, and Junghoon Seo. Implicit stacked autoregressive model for video prediction. *ArXiv*, 2023. doi: 10.48550/arXiv.2303.07849.

Prabhishek Singh, Manoj Diwakar, Achyut Shankar, Raj Shree, and Manoj Kumar. A review on sar image and its despeckling. *Archives of Computational Methods in Engineering*, 28(7):4633–4653, 2021. ISSN 1886-1784. doi: 10.1007/s11831-021-09548-z. URL http://dx.doi.org/10.1007/s11831-021-09548-z.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

V. V. Tikhonov, A. N. Romanov, I. V. Khvostov, T. A. Alekseeva, A. I. Sinitskiy, M. V. Tikhonova, E. A. Sharkov, and N. Yu. Komarova. Analysis of the hydrological regime of the gulf of ob in the freezing period using smos data. *Rossiyskaya Arktika*, 2022. doi: 10.24412/2658-4255-2022-2-44-71.

Kevin Trebing, Tomasz Stanczyk, and Siamak Mehrkanoon. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, 145:178–186, may 2021. ISSN 0167-8655. doi: 10.1016/j.patrec.2021.01.036. URL http://dx.doi.org/10.1016/j.patrec.2021.01.036.

M. V. Tretiakov and A. I. Shiklomanov. Assessment of influences of anthropogenic and climatic changes in the drainage basin on hydrological processes in the gulf of ob. *Water Resources*, 49 (5):820–835, sep 2022. ISSN 1608-344x. doi: 10.1134/s0097807822050165. URL http://dx.doi.org/10.1134/S0097807822050165.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, jan 1990. ISBN 9781611970128. doi: 10.1137/1.9781611970128. URL http://dx.doi.org/10.1137/1.9781611970128.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, ACSSC-03. IEEE, 2003. doi: 10.1109/acssc.2003.1292216. URL http://dx.doi.org/10.1109/ACSSC.2003.1292216.

Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, apr 2004. ISSN 1057-7149. doi: 10.1109/tip.2003.819861. URL http://dx.doi.org/10.1109/TIP.2003.819861.

Weather and Climate. Weather archive. http://www.pogodaiklimat.ru. URL http://www.pogodaiklimat.ru.

John Wallace Weatherly and John E. Walsh. The effects of precipitation and river runoff in a coupled ice-ocean model of the arctic. *Climate Dynamics*, 12:785–798, 1996. doi: 10.1007/s003820050143.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *ICCV*, 2023.

Daniel J. Weiss, Peter M. Atkinson, Samir Bhatt, Bonnie Mappin, Simon I. Hay, and Peter W. Gething. An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:106–118, dec 2014. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2014.10.001. URL http://dx.doi.org/10.1016/j.isprsjprs.2014.10.001.

Da Wu, Xiao Lang, Wengang Mao, Di Zhang, Jinfen Zhang, and Rong Liu. Vae based non-autoregressive transformer model for sea ice concentration forecast. *International Ocean and Polar Engineering Conference*, 06 2022.

Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. Motionrnn: A flexible model for video prediction with spacetime-varying motions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, jun 2021. doi: 10.1109/cvpr46437.2021.01518. URL http://dx.doi.org/10.1109/CVPR46437.2021.01518.

Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris Holmes, Frank Hutter, and Yee Whye Teh. Neural ensemble search for uncertainty estimation and dataset shift, 2022.

Hanqing Zeng, Hanjia Lyu, Diyi Hu, Yinglong Xia, and Jiebo Luo. Mixture of weak and strong experts on graphs, 2023.

Juntang Zhuang, Nicha C Dvornek, Sekhar Tatikonda, and James Duncan. {MALI}: A memory efficient and reverse accurate integrator for neural {ode}s. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=blfSjHeFM_e.

APPENDIX

A DATA DESCRIPTION

TARGET REGION

The Gulf of Ob, located at the mouth of the Ob River in the Arctic Ocean, is the world's longest estuary, stretching approximately 1,000 km between the Gyda and Yamal peninsulas (Lapin et al., 2011). It is relatively shallow, with depths averaging 10 to 12 meters, limiting heavy sea transport.

The Taz Estuary, formed by the Taz River, spans about 330 km from Tazovsky to the Gulf of Ob, with an average width of 25 km. It flows north to south and then bends westward before merging with the Gulf of Ob, contributing to one of the largest estuarine systems in the world.

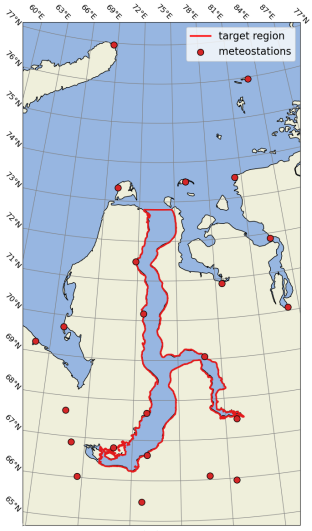
This region is important for sea ice forecasting and research due to its highly variable ice conditions influenced by seasonal changes and river discharge (Osadchiev et al., 2021). It's a sensitive indicator of climate change and has significant economic and strategic value due to its location near major shipping routes and natural resources (Tretiakov & Shiklomanov, 2022). The unique interaction between river outflows and the sea creates distinctive ice patterns, making it a key area for studying sea ice dynamics and improving predictive models(Tikhonov et al., 2022). Additionally, sea ice in this region affects local ecosystems and communities, highlighting the broader impacts of environmental changes on ecology and society.

INPUT FIELDS

Our neural network model utilizes a number of input channels (fields) that come from three sources: Sentinel-1(Sentinel-1), Global Ocean Physics Reanalysis (GLORYS)(GLO), and historical data from meteostations(Weather & Climate) (see Figure 5 for detailed information). Sentinel-1 SAR images are interpolated conservatively to match the input resolution (1 km), GLORYS fields are interpolated bilinearly, data from meteostations is interpolated between discrete points (where the meteostations are located) using RBF interpolation method with thin plate splines (Wahba, 1990). The details on resulting channels and preprocessing for input data are described in Table 5.

Table 5: Description of input channels. GLORYS channels are interpolated bilinearly. Meteodata is interpolated using radial basis function interpolation.

Source	Scale	Channel	Normalization
Sentinel-1	1 km	SAR HV	$U(0,1)$
		SAR HH	$U(0,1)$
GLORYS	5 km	Bottom Temperature	$U(-1,1)$
		Mixed Layer Thickness	$U(-1,1)$
		Surface Salinity	$U(-1,1)$
		Surface Temperature	$U(-1,1)$
		Sea Ice Velocity (u)	$N(0,1)$
		Sea Ice Velocity (v)	$N(0,1)$
		Sea Height	$N(0,1)$
Meteostations	–	Relative Humidity	$U(0,1)$
		Air Pressure	$N(0,1)$
		Air Temperature	$N(0,1)$
		Wind Velocity (u)	$N(0,1)$
		Wind Velocity (v)	$N(0,1)$



(a)

Name	Lat	Lon
New Port	67.68	72.88
Seyakha	70.17	72.52
Tambay	71.48	71.82
Popov Island	73.33	70.05
Nyda	66.63	72.93
Yar-Sale	66.80	70.83
Beloyarsk	66.87	68.13
Laborovaya	67.64	67.56
Poluy	66.03	68.68
Nadym	65.47	72.67
Tazovsky	67.47	78.73
Antipayuta	69.08	76.85
New Urengoy	66.10	76.78
Urengoy	65.95	78.40
Vilkitsky Island	73.50	76.00
Dixon Island	73.52	80.40
Marresal	69.71	66.80
Ust-Kara	69.25	64.93
Karaul	70.08	83.17
Sopochnaya Karga	71.87	82.70
Izvestia Islands CEC	75.95	82.95
Cape Zhelaniya	76.95	68.55
Gyda	70.88	78.47

(b)

Figure 5: The map (a) and coordinates (b) of meteorological stations used, along with the target region outlined in red. Available sea surface area is 120,559 km^2. The area of interest is 51,262 km^2.

SAR ESTIMATES SEA-ICE CONDITIONS

In comparison to other potential target variables, such as GLORYS reanalysis, which lacks quality in the Gulf and which is mostly uncorrelated with other sources (see Figure 7), GLORYS operative analysis and forecasts, which lack historical records essential for data-driven approaches, and AMSR (Ludwig et al., 2020), which is partly dependent on cloud conditions and seasons, Sentinel-1 SAR imagery emerges as a superior choice for high-resolution sea ice forecasting models.

While the direct comparison between SAR and calculated sea ice concentrations is not strictly fair, the techniques of retrieval and mapping sea ice conditions from SAR imagery are well-known. Sentinel-1 C band consists of four polarizations, for the purpose of forecasting ice conditions we utilize colorized HV polarization (Sentinel Hub). Figure 6 shows the comparison of monthly-averaged sea ice concentrations from several sources.

B METRICS

In our research we utilize two type of metrics. First, we use the common computer vision ones: the mean squared error (MSE) also known as L2-distance; the structural similarity index measure (SSIM), a metric used to assess the human-perceived quality of digital images and videos (Wang et al., 2004), predominantly used in computer vision; and the multi-scale structural similarity index measure (MS-SSIM), which extends the concept of the SSIM by evaluating image quality at various

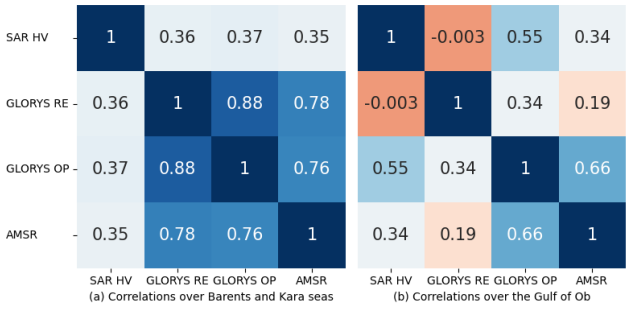
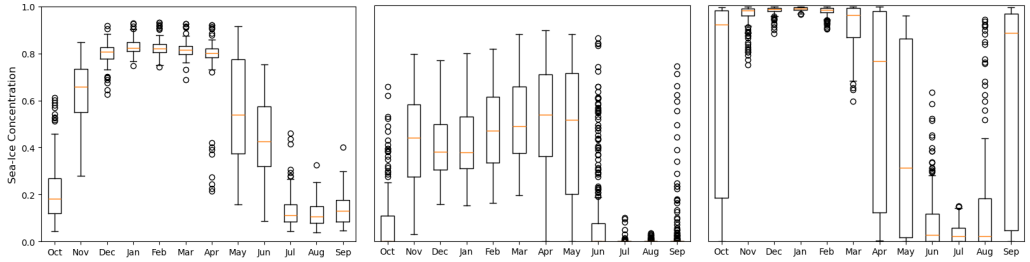


Figure 6: Mean cell-wise sea ice concentration correlation between several data sources: Sentinel-1 SAR (Sentinel-1), GLORYS operative and reanalysis (GLO), and AMSR (Ludwig et al., 2020)



(a) SAR-estimated SIC (b) GLORYS Reanalysis SIC (c) GLORYS Operative SIC

Figure 7: Box and whisker plots of SIC data distribution in the region in GLORYS and Estimated from SAR-images for different months from all available range of time, aggregated over target region. The box extends from the 25th percentile to the 75th percentile; whiskers extend the box by 1.5x of its length. The orange line is the median (50th percentile); SAR-estimated SIC is a normalized mean absolute value of SAR signal with dropped frames with more than 50% missing values.

scales (Wang et al., 2003). The MS-SSIM approach uses the fact that the human eye perceives picture quality differently across varying resolutions, making it a more comprehensive metric for assessing the perceived quality of digital images and videos. Second, we use a geophysical metric specific for sea ice condition analysis and forecast: the integrated ice edge error at level c (IIEE at c), a metric of similarity between ice sheets, where ice edges are chosen at the certain level of concentration c measured in percents (Goessling et al., 2016):

$$IIEE@c = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{S}\sum_{h,w}[(y_i > c) \neq (\hat{y}_i > c)]\, dS_{hw}, \quad S = \sum_{h,w} dS_{hw} \tag{4}$$

where y_i and $\hat{y}_i$ are linearly normalized into range $[0, 100]$. Usually parameter c is set to 15%, however we can not assume a linear relationship between ice concentration and SAR images, thus we will exploit several values for c.

C Optical flow estimation for sea ice

We argue that a fundamental challenge with modern machine learning models is their inability to replicate the complex mechanics of sea ice in coastal regions. The poor performance in capturing fine-scale ice mechanics is not unique to any one method but is a common issue across various approaches at the target resolution. For instance, neither modern sea ice motion vectors (GLO; Noriaki et al., 2013) nor optical flow estimation methods (Farnebäck, 2003; Weinzaepfel et al., 2023; Sun et al., 2018) are well-suited for high-resolution ice velocity estimation. This degradation in quality when transitioning to higher resolutions is illustrated in Figure 8. Moreover, deep learning methods for optical flow estimation may be overfitted on common images and lack the generalization needed for sea ice SAR imagery. Consequently, motion information is scarcely useful for predictions in the region of interest.

Low quality of optical flows might be caused by high homogenity of ice-sheet surface and stochastic local dynamics on 1-day scale. For similar reasons one can expect state-of-the-art models on video-prediction task to fail on ice-dynamics forecasting, as their architectures sometimes are based on optical flow estimation and prediction, and they assume the simple mechanical and deterministic dynamics.

Resolution:	1 km	2 km	4 km	8 km	16 km	32 km
Persistence	7.0	7.2	7.2	7.4	7.1	7.5
Glorys Operative	8.6	8.8	8.2	8.4	6.9	**5.6**
AMSR JAXA SIM_R	7.1	6.9	6.6	**6.4**	**6.3**	5.8
Farneback	**6.7**	**6.6**	**6.4**	**6.4**	6.5	5.9
CrocoFlow	6.8	6.7	6.5	**6.4**	6.5	5.9
PWC-Net	6.9	6.8	6.5	**6.4**	7.0	8.2

Figure 8: Mean Squared Error (MSE) ($\times 10^{-3}$) between next-day images and previous-day images, warped using estimated flow from following sources: GLORYS Operative model (25 km resolution), AMSR JAXA SIM-R (50 km resolution), and several Optical Flow models, such as the algorithmic Farneback method and state-of-the-art neural networks. The best MSE values for each resolution are highlighted in bold.

D Ablation Study

This section contains ablation studies for crucial parts of training and prediction pipelines: filtration of SAR-imagery artifacts (Table 6), proper augmentations to leverage unbalance and lack of data (Table 7), and the usage of confidence-based model selection and ensembles (Tables 8, 9).

Ensembles usually provide minor improvements except for IIEE@15 metric. However, confidence based model selection suppresses the advantages of ensembles. The usage of model selection (depicted at Table 2) increases MSE and IIEE@75 by 12%.

Table 6: Summary of the metrics obtained by testing individual models without data preprocessing. Raw data have high noise-to-signal ratio due to thermal artifacts. These artifacts simultaneously provide huge bias in metrics and corrupt loss function making the models learn filtration and smoothing rather than forecasting sea ice dynamics.

Model	MSE	1 - SSIM	1 - MS-SSIM	IIEE@15	IIEE@30	IIEE@50	IIEE@75
	($\times 10^{-3}$)			($\times 10^{-2}$)			
Persistence	18.5	9.6	6.8	17.3	12.5	10.2	8.5
Linear	15.7	8.9	6.2	17.6	12.9	10.0	8.4
DMVFN	16.7	8.5	6.2	17.2	12.4	9.9	8.1
IAM4VP	16.8	9.7	6.6	18.5	15.5	11.5	10.4
Neural ODE	13.7	8.5	5.8	17.0	12.4	9.9	7.8
MotionRNN	12.7	8.1	5.4	16.2	11.9	9.2	7.6
Vid-ODE	12.2	7.8	5.4	16.5	11.4	8.7	7.1
UNet	13.0	7.6	5.4	15.7	11.4	9.1	7.4
rUNet	13.6	7.8	5.5	15.7	12.0	9.3	7.6

Table 7: Ablation studies for the augmentations for the best performing model (rUNet). Geometric augmentations are shifts and rotations (treating input as an image). The physical augmentations are modifications of geometrical ones with corresponding transform (rotations and flips) of physical fields (currents and winds). "Proposed" states for superposition of Physical and Frameout augmentations.

Augmentation	MSE	1 - SSIM	1 - MS-SSIM	IIEE@15	IIEE@30	IIEE@50	IIEE@75
	($\times 10^{-3}$)			($\times 10^{-2}$)			
None	8.9	9.2	4.9	11.9	9.4	10.8	7.0
Geometric	8.7	9.0	4.8	12.9	10.4	10.9	6.6
Physical	7.8	8.4	4.6	10.1	9.1	10.0	6.1
Proposed	7.6	8.3	4.6	10.0	9.0	9.8	6.0

Table 8: Summary of test metrics for individual models with proposed preprocessing and augmentations.

Model	MSE	1 - SSIM	1 - MS-SSIM	IIEE@15	IIEE@30	IIEE@50	IIEE@75
	($\times 10^{-3}$)			($\times 10^{-2}$)			
Persistence	11.2	9.8	5.6	11.5	10.4	11.0	7.3
Linear	9.9	9.1	5.2	14.2	9.6	11.0	8.0
DMVFN	10.0	8.8	5.1	11.7	10.2	10.8	6.9
IAM4VP	9.5	10.6	5.6	14.7	10.6	11.4	8.1
Neural ODE	8.6	9.3	4.9	12.1	10.1	10.7	6.8
MotionRNN	8.0	9.0	4.7	11.4	9.3	10.3	6.5
Vid-ODE	7.7	8.6	4.7	12.2	9.2	9.6	6.0
UNet	8.3	8.2	4.6	12.1	9.5	9.9	6.5
rUNet	7.6	8.3	4.6	10.0	9.0	9.8	6.0

Table 9: Summary of test metrics for ensembles. "rUNet x3" stands for mean forecast of 3 retrained versions of rUNet. "Best 4" stands for mean of MotionRNN, Vid-ODE, UNet, and rUNet predictions.

Ensemble	MSE	1 - SSIM	1 - MS-SSIM	IIEE@15	IIEE@30	IIEE@50	IIEE@75
	($\times 10^{-3}$)			($\times 10^{-2}$)			
rUNet x3	7.1	8.3	4.5	11.2	8.8	9.2	6.1
Best 4	7.1	8.2	4.4	11.2	8.7	9.4	6.1

E SUPPLEMENTARY FIGURES

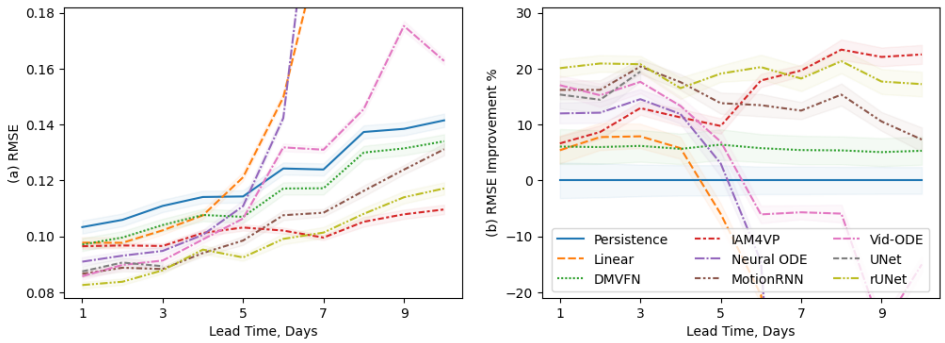


Figure 9: (a) RMSE and (b) its percentage improvement over persistence baseline for each extended lead time in days over the test subset.

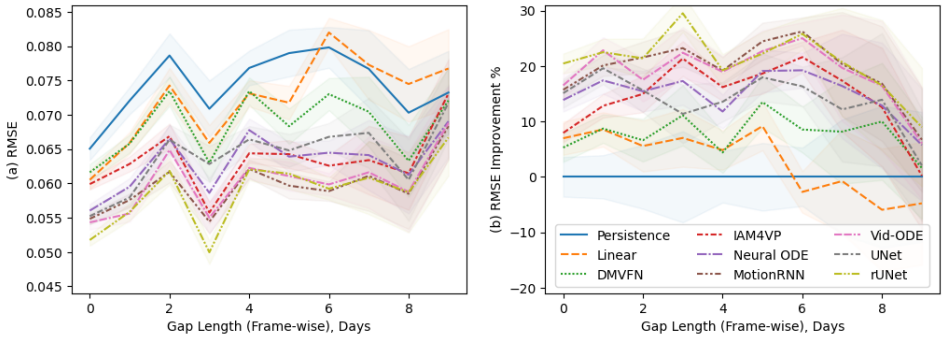


Figure 10: (a) RMSE and (b) its percentage improvement over persistence baseline in dependence of preceding SAR gap length.

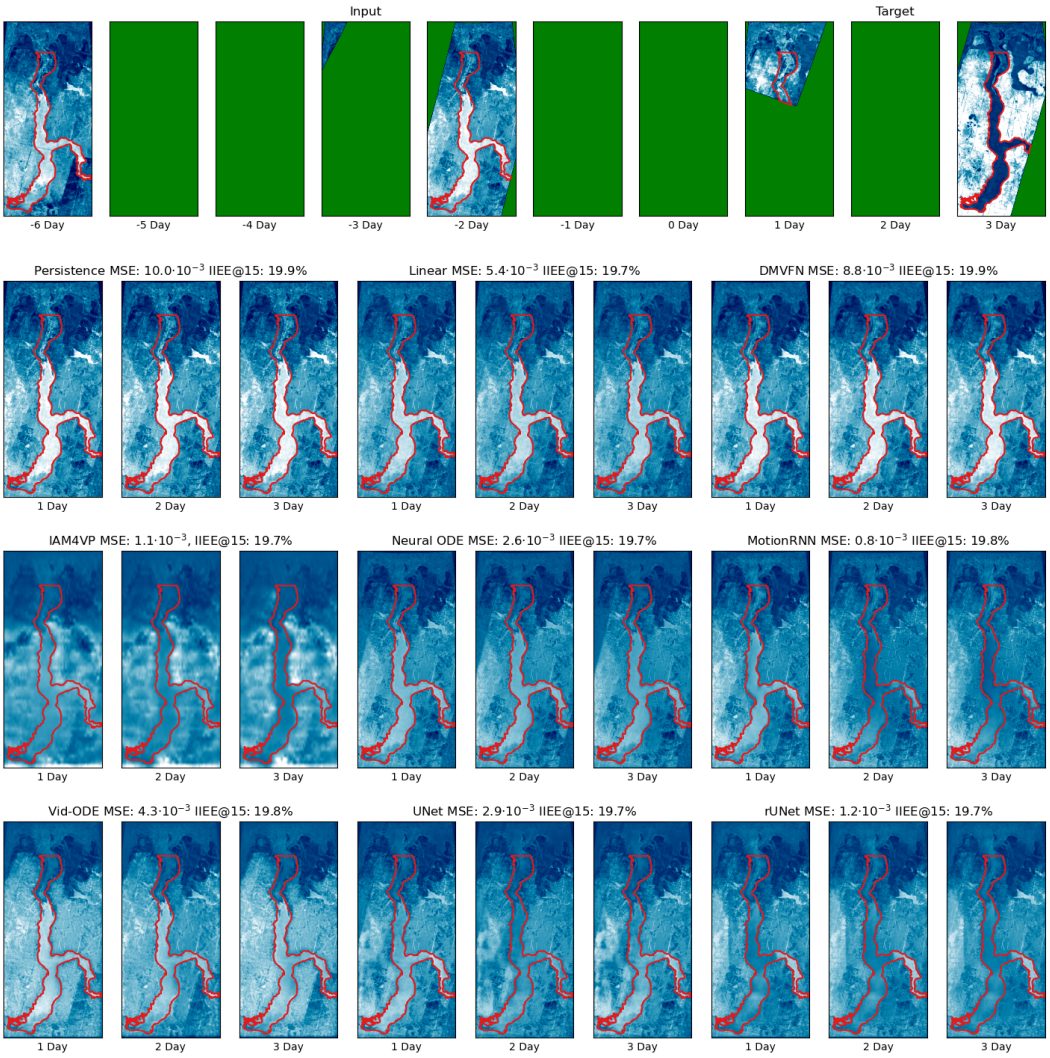


Figure 11: The example of forecasts. Timestamps represent shifts from the 25-05-2023. The target region is outlined with a red line. The missed data in an input and a target sequences is represented by green color.

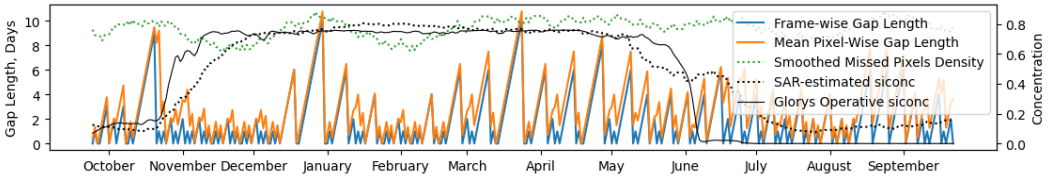


Figure 12: Distance to the nearest valid data frame-wise (blue) and mean value pixel-wise (orange); concentration of missing values smoothed with half-month-wide rolling window; operative glorys sea-ice concentration; and mean SAR-value as an estimation of sea-ice concentration. All curves are evaluated over test subset.

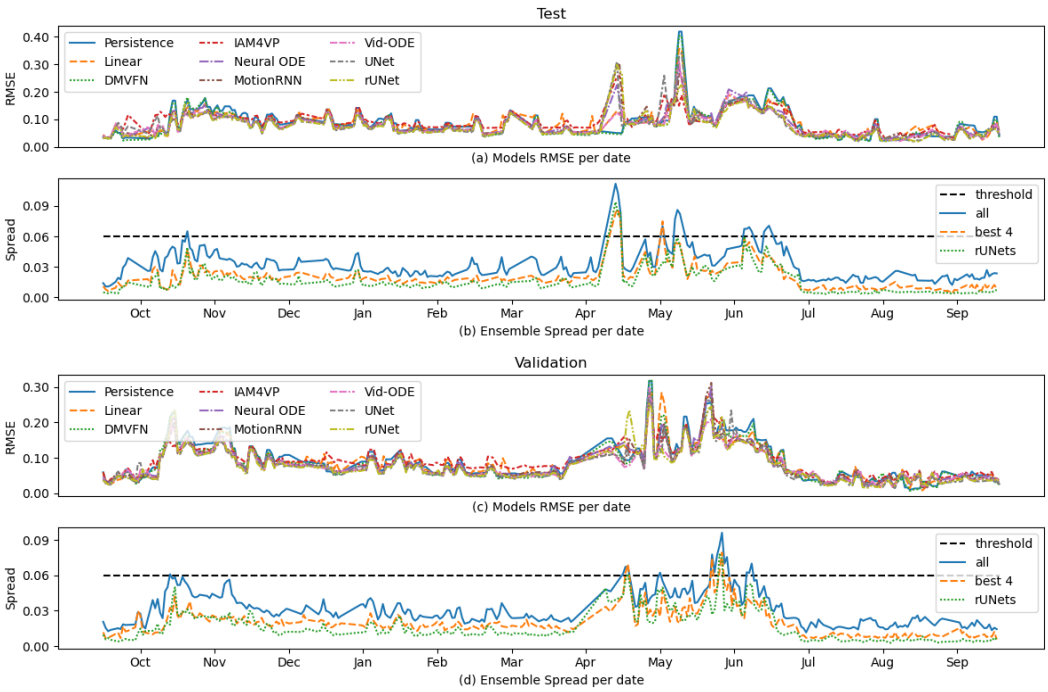


Figure 13: RMSE timelines for all the models over test (a) and validation (c) subsets, ensemble spreads over test (b) and validation (d) subsets, where 'best 4' stands for MotionRNN, Vid-ODE, UNet, and rUNet, 'rUNets' stands for 3 different initializations of rUNet model. Threshold is tuned on validation subset for consequent use in confidence-based model selection during testing.

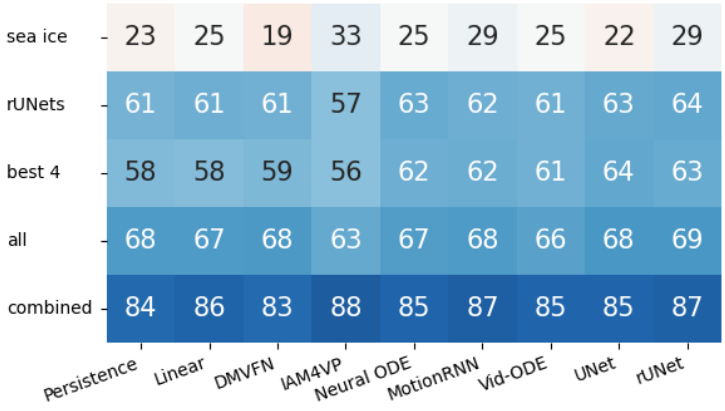


Figure 14: Correlation (in percents) between models RMSE (with confidence-based model selection) and several features: sea ice concentration, ensemble spread, and their learned linear combination.