Understanding the Limitations of Deep Transformer Models for Sea Ice Forecasting

Julia Borisova[1 [0009−0003−6680−0940]], Andrey Kuznetsov[1 [0009−0009−2299−8186]], Gleb Solovev[1 [0009−0005−5479−2482]], and Nikolay O. Nikitin[1 [0000−0002−6839−9957]]

ITMO University, St. Petersburg, 197101, Russia

Abstract. It would not be an exaggeration to say that we live in the era of transformers. Due to the great results of generative models for video prediction, spatio-temporal data of various kinds are usually treated as video-like sequences - and this is a good assumption for many problems. However, we want to argue that transformer-based prediction is not the best option for some spatio-temporal cases with regular grid and strong periodicity (since most discussions about the limitations of transformer applicability focus only on time series).
In the paper, we considered the task of sea ice forecasting and analyzed two transformer-based architectures (TimeSformer and SwinLSTM) against the proposed baseline - a lightweight convolutional network with different setups of convolutional layers (2D and 3D). Experiments for long-term forecasting of Arctic seas show that transformers do not reproduce the annual dynamics of sea ice. At the same time, the CNN-based solutions allow to outperform the existing state-of-the-art numerical (SEAS5) and data-driven (IceNet) forecasts, with a quality improvement of up to 30% in the mean absolute error and up to 10% in the structural similarity index. A similar experiment is provided for the synthetic example of video data. Due to the analysis of the obtained results, this problem is caused by the nature of the model and the data and can be faced in many scientific and industrial tasks outside sea ice.
Code and supplementary materials for this research are available on GitHub: `https://github.com/ITMO-NSS-team/sea_ice_transformers`.

Keywords: sea ice concentration · transformers · CNN · spatio-temporal data · long-term forecasting.

1 Introduction

The discussion on the limitations of the applicability of transformers is widely presented in the literature [25], especially for the forecasting of time series. However, for the broad class of spatio-temporal tasks (e.g. video prediction), transformers are considered as the basis of almost any state-of-the-art model. One of the most challenging spatio-temporal problems for AI is sea ice forecasting [12].

Since spatio-temporal data have similarities with a video sequence and its forecasting can be considered as a video prediction task, it looks promising to apply more complex architectures with the attention mechanism - transformers

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

- that have proven successful for processing video [14]. While the application of transformers for long-term, high-resolution forecasting of the entire area of the Arctic Ocean is challenging (due to the amount of computational resources required to train a model), it appears suitable for the task of regional modelling of ice dynamics in specific water areas, which is also quite important [5].

We found that *self-designed baseline models based on simple convolutional architecture significantly outperform deep transformers for long-term regional sea ice forecasting* for all regions considered. Similar problems with the applicability of transformers are widely discussed for univariate and multivariate time series forecasting [25]. However, for tasks similar to video prediction, the limits of the applicability of transformers are still not well discussed. Therefore, we conducted a detailed investigation using different models and setups.

In the experimental part of the paper, we provide a comparison of four approaches to predictive modelling of ice conditions. Two transformer-based decisions: TimeSformer [3], which adapts the architecture for sequence prediction, and SwinLSTM [20]. Shallow two- and three-dimensional convolutional networks have been proposed as strong baselines. The experimental setup includes regional prediction of ice concentration for five Arctic seas.

Experiments show the inability of transformers to reproduce the annual dynamics of ice concentration due to incorrect periodicity components. At the same time, CNN-based baselines provide adequate results for long-term forecasting. We compare them with two state-of-the-art (SOTA) solutions for sea ice forecasting (physics-based system SEAS5 [10] and neural ensemble-based model IceNet [1]). *The improvement of the CNN-based baseline over domain-specific SOTA is up to 30% in mean absolute error, up to 10% in structural similarity index and up to 6% in ice edge reconstruction accuracy.*

We can conclude that it is important to extend the existing benchmarks for video prediction by novel tasks with strong periodic component to represent the limits of transformers for a wide class of real-world "periodic" tasks (from remote sensing to the analysis of production systems [6]). We have also provided a synthetic example of periodic data for video forecasting task as an empirical proof that the considered problem is not specific only to sea ice.

2 Related works

Sea ice forecasts can be obtained using global physical models based on systems of differential equations have been developed to simulate ice conditions.However, despite the scale of the system, the simulation is global and roughly reproduces local processes, which is a serious limitation of its applicability For this reason, regional physical models such as SI^3 are widely used for real-world tasks.

The accumulation of remote sensing data has made it possible to use fully data-driven models to forecast ice conditions. The task is known as spatio-temporal forecasting, so image-based deep learning models are actively used. Disadvantages of most deep learning models for forecasting ice conditions include a short forecasting horizon (several days or weeks [8]), while long-term

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

forecasting (>3 months) is especially important for planning industrial work for the next seasons.

Convolutional Neural Networks (CNN) [13] are widely used in sea ice modelling. The most popular architecture is the U-net [1,8]. There are examples of the use of U-net models for long-term forecasts up to one year [11]. However, the limiting factor for the use of such solutions is the need for a large amount of additional input data on the atmosphere (temperature, pressure, solar radiation, etc.) for training and inference. To improve the quality of forecasts, simple models are often combined into ensembles, which allows a probabilistic modelling component to be introduced, taking into account the confidence of each of the ensemble models [1].

Since the spatio-temporal data is similar to a video sequence, video prediction methods can be applied to sea ice concentration forecasting. The first group of methods are recurrent networks. There are many architectures from ConvLSTM [18] to the more recent (e.g. CrevNet [24]), which proposes a CNN-based recurrent network for learning spatio-temporal dependencies. The PhyDNet [9] model introduces physical knowledge into a CNN-based model to improve the quality of prediction. For video prediction, these models perform reasonably well due to their ability to account for spatial and temporal dependencies.

The Transformer [22] architecture has also been widely applied to video processing. The ViT [7] model was the first to use Transformers directly for image classification and achieved impressive results. However, the performance of the ViT model is highly dependent on the size of the training sample. There is also a promising SwinLSTM [20] model for video prediction based on Swin Transformer [14] blocks with a simplified LSTM. This model performs well in analysing temporal and spatial dependencies in video files.

Limitations of transformers is a topic that is widely discussed in the literature [25]. There is even a repository *Transformers And LLM Are What You Dont Need* [1], which contains the examples where simple models overcome deep transformers for different tasks. However, the papers in this repository focus on time series data (both univariate and multivariate). Even if the data has a spatiotemporal nature (e.g. the task of predicting traffic at different spatial points , it is not represented as a regular grid and solved by other methods (e.g. graph neural networks). Thus, the limitation of the applicability of transformers to video-like sequences is still an under-discussed topic.

3 Problem statement

We consider the problem of regional sea ice concentration prediction not only from a domain-specific point of view. In this paper, we discuss the applicability of state-of-the-art computer vision models to the data with specific properties that are characteristic of the environmental case considered.

[1] `https://github.com/valeman/Transformers_And_LLM_Are_What_You_Dont_Need`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

3.1 Nature of the data and models.

The main difference between metocean forecasting and conventional spatio-temporal forecasting is the different nature of the data. First, the dynamic processes in environmental systems are multiscale and non-stationary. In addition, they contain an irremovable stochastic component.

However, state-of-the-art computational methods still perform well on a large part of natural systems forecasting tasks - for example, the transformer-based basic model can outperform both state-of-the-art classical simulation tools and specialised deep learning models for weather forecasting [4]. So what is the problem? Why can we not apply state-of-the-art CV tools directly to sea ice data? What is so special about this?

One issue is the non-differentiable nature of sea ice data - it is not a smooth field, but data with a clear distinction between concentrated ice and clean water (the so-called "ice edge"). In addition, sea ice has very complex periodic patterns (e.g. annual periodicity for sea ice), the reproduction of which is crucial for current forecasting.

In this paper we aimed to prove or reject the **hypothesis**: *the practical applicability of regular-grid transformer-based models for spatio-temporal data with specific periodic properties is very limited*. We use regional sea ice concentration prediction as a real-world case study to empirically confirm it.

The theoretical basis of this hypothesis is as follows: artificial neural networks are combinations of several simple mathematical functions that implement more complex functions from one real data value to another. The spaces of multivariate functions that can be implemented by a network are determined by the structure of the network and its parameters. Ice concentration data is non-linear, it is a time series with pronounced periodicity. Since neural network architectures based on transformers have a linear nature [17], the main hope for improving the quality of prediction is achieved through a large number of parameters.

It is known that adding the ReLU activation function allows to increase the efficiency of networks on linear layers by transforming the model architecture [23]. Therefore, in the process of adapting the transformers to the task, ReLU activation functions were added to the architecture to improve the quality of data approximation based on a large number of parameters.

3.2 Benchmarks for spatio-temporal tasks.

As the task of spatio-temporal prediction is not new, there are many well-known open benchmarks against which the model can be compared. For example, the OpenSTL[2] [19] benchmark for spatio-temporal predictive learning covers several tasks (including weather prediction from WeatherBench [16]). However, tasks similar to sea ice forecasting are not included in these benchmarks. For this reason, we cannot base our experimental setup on existing benchmarks and prepare our own dataset.

[2] https://github.com/chengtan9907/OpenSTL

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

4 Proposed approaches

We propose a strong baseline for the task of sea ice prediction based on a convolutional architecture. As typical examples of transformer models, we choose TimeSformer and SwinLSTM. The mean absolute error was used as the loss function for all models. The technical details of the model implementations and their adaptation to the sea ice forecasting task are given below.

4.1 Baseline

CNN-2D was implemented as a CNN with an encoder-decoder architecture. It consists of 5 convolutional 2D layers with ReLU activation function and its transposed mirror. The values in the input images range from 0 to 1 due to the nature of the ice concentration data. As input data, the model receives a multichannel image with the history of the parameter; the output of the model is a multichannel image with a prediction n steps ahead.

The training sample was formed by a sliding window along the space-time series. The scheme illustrating the dataset formation is shown in Figure 1. This approach allows the starting point of the model to be varied and a forecast to start on any day of the year. This is important for applying the model to real industrial problems as the forecast can be based on the most recent data.

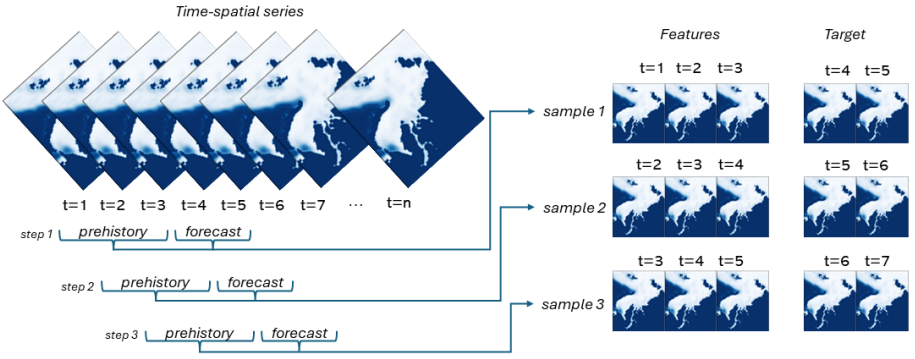

Fig. 1. The preprocessing of training set of sea ice forecasting

Models trained with the L1 loss function tend to produce grain artefacts during inference. To solve this problem and make the model lighter, we reduced the spatial resolution of the input images (by a factor of 2).

Baseline CNN-3D uses a time component sensitive CNN encoder-decoder architecture with 3D convolution. Each of the encoder and decoder parts consists of 2 layers, forming a symmetric structure. Otherwise, the architecture and training process are identical to the previous model.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

The model has fewer layers and parameters than the baseline because 3d convolution is asymptotically more complex. This increased complexity means that convolution operations with a 3d kernel can be more time-consuming than those with a 2d kernel, even when the number of parameters is reduced. For example, with almost the same number of parameters for 2D and 3D convolutions (2234 and 2529 estimated with software), their total number of multi-adds is 42.83 million and 216.32 million respectively. The number of parameters for TimeSformer is 33 million and 20 million for SwinLSTM. The impact of such an increase on the runtime is shown in the Table 2.

The third dimension of the kernel acts as a temporal dimension to the input, which consists of 2D images over time. By tuning this third dimension, the model can effectively extract temporal components such as seasonality or trends, thereby improving its ability to detect and predict time-dependent patterns in the data. For these reasons, 52 was chosen as the third dimension of the kernel, corresponding to one year of prediction, with each time step representing one week. Choosing a higher frequency or increasing the number of layers becomes challenging because with the input size halved and a history of 104 timesteps, the feature maps can degenerate to zero after convolution.

4.2 Transformers

The dynamics of ice concentration changes can be represented as spatio-temporal data closest to the video of ice melt and ice intrusion. In this video series, not only neighbouring images are linked, but there can also be a link between images related by seasonality. For example, the data for January of each year are linked, and taking this into account it is possible to predict February more effectively. An appropriate attention mechanism can be used to identify and take account of this relationship. Although the original Transformer was developed for NLP (natural language processing) tasks, there are now solutions for processing images, such as ViT [2] and Swin Transformer [14]. For comparison with the proposed baseline solutions, two transformer-based models were applied to the ice concentration forecasting problem: TimeSformer [3] and SwinLSTM [20].

TimeSformer. Processing frames alone is not enough to create an effective approach to sea ice forecasting. ViViT [2]and TimeSformer can provide a more in-depth method for processing data such as video. These models are designed for the task of video series classification and are encoder-only models.

TimeSformer implements the Divided Space-Time attention mechanism, which we believe is the promising basis for the sea ice prediction task. Thus, our experiments with transformers are based on the developments of TimeSformer.

The model architecture had to be refined because the original TimeSformer was designed for video classification, and the task at hand requires the prediction of ice concentration changes over multiple frames. To this end, the transformer head responsible for classification was replaced by a convolutional decoder. This decoder translates the hidden state of the output data from the TimeSformer

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

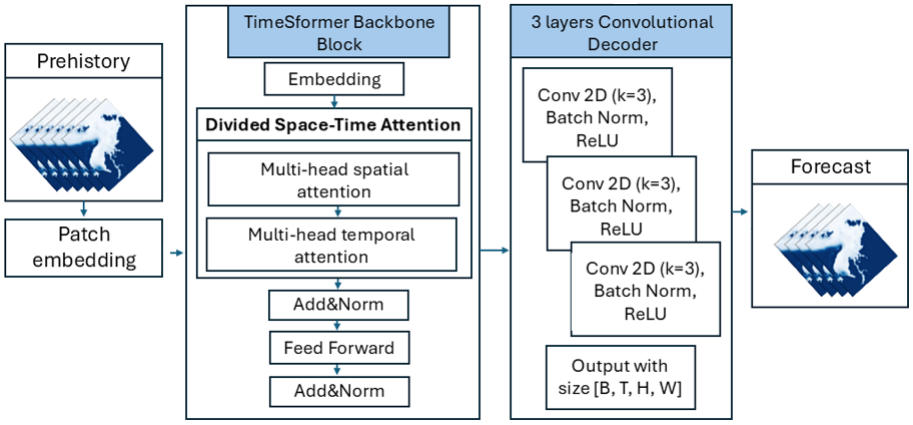

Fig. 2. Application of the TimeSformer to the long-term sea ice forecasting

backbone into the $[B,T,H,W]$ [3] dimension of the sea ice data set. Three convolution layers with ReLU activation function and BatchNorm normalization were used to unlock the decoder. A schematic illustration of the transformer for ice concentration forecasting is shown in Figure 2.

Two NVIDIA Tesla P100 GPUs were used to train the model, and the time spent is shown in the Table 2. The total number of epochs was set at 120, in accordance with the estimates used by the authors of the architecture (the original work trains the model for 15 epochs). We also performed the additional experiments and made sure that increasing the number of epochs did not improve the results. To ensure the adequacy of the chosen number of epochs, convergence curves were constructed for the training and test samples (presented in the supplementary material).

SwinLSTM. This approach has performed well for video sequence prediction on well-known datasets such as Moving MNIST, TaxiBJ, Human3.6m, and KTH. The SwinLSTM architecture is based on Swin Transformer blocks and the simplified LSTM. This approach is successful in extracting spatio-temporal representations.

This model was used in both SwinLSTM-D and SwinLSTM-B without significant architectural changes. The only change in our approach was to change the resolution of the input data. Since the original SwinLSTM used data with resolutions of 32x32, 64x64 and 128x128, it is expected that this model can be successfully applied to our data resolutions. The learning process, optimizer, loss function and learning rate values have not been changed. The only limitation of this model is the frame prediction range. The authors of the paper conducted experiments to predict from 4 to 14 consecutive frames. Our experiment requires

[3] (B - batch, T - time, H - height, W - width).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

the prediction of 52 frames (52 weeks in a year). The time taken to train the model in this statement for the different test areas shown in the Table 2. For the 6-frame (monthly temporal resolution) prediction training took 13 ± 3 hours for 90 epochs. However, this setup does not allow the intra-month dynamics of sea ice to be represented.

5 Experiments studies

The experimental setup in the paper is focused on comparing the performance of the proposed baseline model and transformer architectures in the sea ice forecasting task.

We use the OSI SAF Global Sea Ice Concentration [21] product as training data. The spatial resolution of the images is reduced to 14 km. To test the generalisability of the developed models for different water areas, five Arctic seas were selected as test areas. The spatial position of each sea is shown in the supplementary materials.

The forecast horizon for the predictive models was set at one year ahead in order to produce long-term forecasts. For inference, the pre-history length was set to two years. The time resolution of the series was set to 7 days. Models were trained over the period 1979 to 2020 years. Dates from 01/01/2020 to 31/12/2023 are used as a test set.

In order to compare the predictive capabilities of the models, forecasts were made on the test sample starting on 1 January of each year. The quality metrics chosen were the mean absolute error (MAE) for each prediction step and the structural similarity index (SSIM). Ice edge product can be computed with ice concentration through binarization. The choice of threshold is due to studies [1] as a marker for the presence of ice in remote sensing data. Binary accuracy on predicted ice edge was calculated to indicate the quality of thick ice position prediction. Averaged metrics for the test sample are presented in Table 1. For convenience, expanded tables with metrics averaged by quarters of each test year are presented in the supplementary materials.

Architectures based on 2D and 3D convolutional layers differ significantly in the complexity of the operations performed. Measurements of the time taken to train 1000 epochs on NVIDIA GeForce RTX 4080 for each of the architectures, depending on the size of the input area, are presented in the Table 2.

According to the metrics, the TimeSformer has a lower quality compared to simpler models. To understand which period of the year makes the largest contribution to the average error, the course of the metric over the year is shown in the Figure 3 for the Kara Sea test area. Vertical lines mark the beginning of each year. There is a pattern in the plot of the TimeSformer error that differs from other models - the error increases significantly in the summer months.

To understand the reason for the increase in TimeSformer error in the summer period, we look at the ice concentration maps for each prediction time step. Example of each model prediction in July compared to the ground truth map shown in Figure 4.We also plot the time series of each prediction at one point to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

Table 1. Quality metrics for implemented models (averaged over 2020-2023), forecast horizon - 1 year (bold are the best)

Metric	Mean Absolute Error (MAE)			Structural Similarity Index (SSIM)			Accuracy (0.2 threshold)		
Model	2D CNN	3D CNN	Time Sformer	2D CNN	3D CNN	Time Sformer	2D CNN	3D CNN	Time Sformer
Kara Sea	**0.080**	0.082	0.111	**0.673**	0.655	0.530	**0.929**	0.928	0.892
Barents Sea	0.065	**0.061**	0.134	**0.679**	0.670	0.482	0.935	**0.937**	0.839
Laptev Sea	**0.075**	0.079	0.161	**0.720**	0.700	0.591	0.933	**0.934**	0.875
East-Siberian Sea	0.087	**0.081**	0.176	**0.710**	0.705	0.688	0.923	**0.931**	0.870
Chukchi Sea	0.084	**0.083**	0.153	0.696	**0.700**	0.567	**0.936**	0.935	0.899

make sure that the pattern of error does not change over the years. As we can see from the maps and plot, TimeSformer does not predict ice melt during the summer period correctly.

TimeSformer. Results are related to the way data are transformed when fed into the spatial attention and temporal attention blocks. In the original work this solution performs well for the video series classification task, however, for predicting ice concentration this solution is not optimal. Perhaps, this solution requires modernization of the input data patching, for example, using 3D convolution as implemented in the ViViT model, as well as applying a different approach in the attention blocks. However, these changes may lead to higher computational complexity, which will eventually require much more computational resources to achieve the quality of models based on 2D convolutions.

SwinLSTM. This model demonstrated low efficiency in forecasting of long sequences. In the considered formulation of the problem of predicting one year from two years of prehistory, SwinLSTM could not achieve the results of TimeSformer. This model predicts all 52 weeks with one coarse value of ice concentration. This prediction is even worse than predicting each week with the average ice concentration for the whole year. In the original experiments of SwinLSTM developers, this model did not predict more than 14 frames. Therefore, the failure in predicting 52 frames of ice concentration is not so surprising.

Comparison with SOTA for sea ice. To evaluate the absolute values of the errors of the implemented models, we compare them with the SOTA solution SEAS5 forecast system. SEAS5, ECMWF's fifth generation seasonal forecasting

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

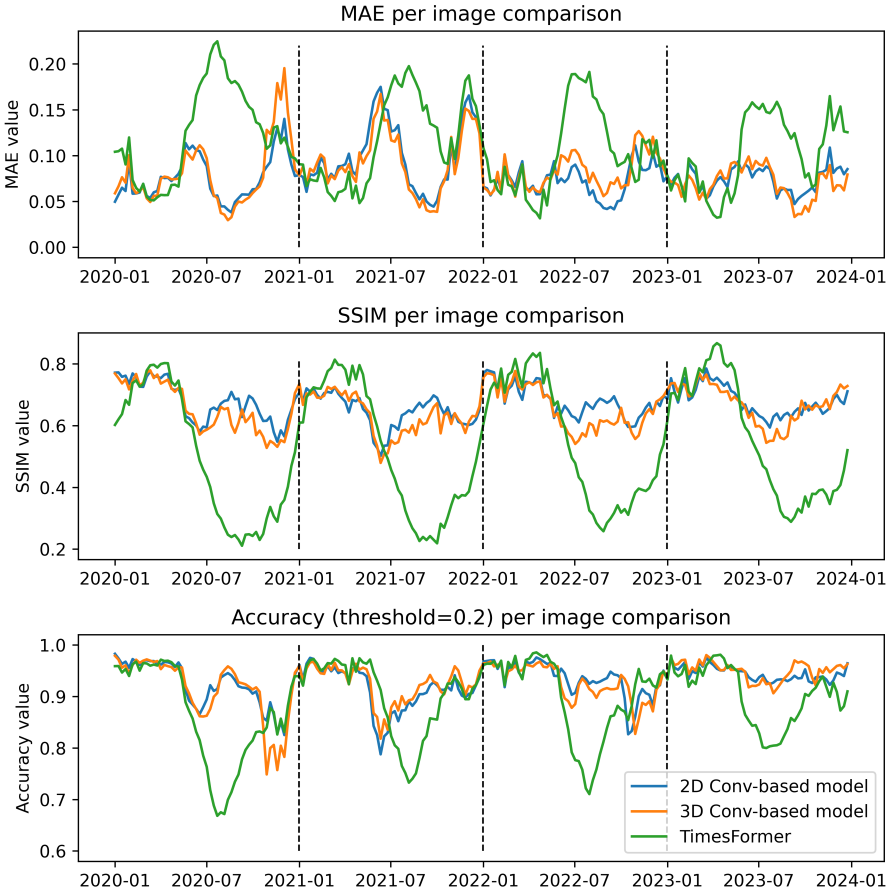

Fig. 3. Metrics for each time step of prediction for Kara sea

system, is physics-based and uses systems of differential equations. It provides a global Arctic forecast 7 months ahead and includes 51 ensemble elements. Due to differences in forecast horizons, in the generalised Table 3 the forecast of all models is limited to the SEAS5 horizon, detailed tables can be found in the supplementary material.

To assess the quality of ice edge prediction, IceNet, a forecasting system based on an ensemble of neural networks, was chosen as a data-driven SOTA. IceNet consists of 25 ensemble members, each of which is a U-net architecture model. Eleven climate and ice cover variables are used as input parameters. As the solution provides a monthly probabilistic forecast for 6 months, we used a confidence threshold of 0.8 for the probabilistic model. In the generalized Table 3 for each of the seas the forecast is limited to a 6 month horizon, a detailed table can be found in the supplementary.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

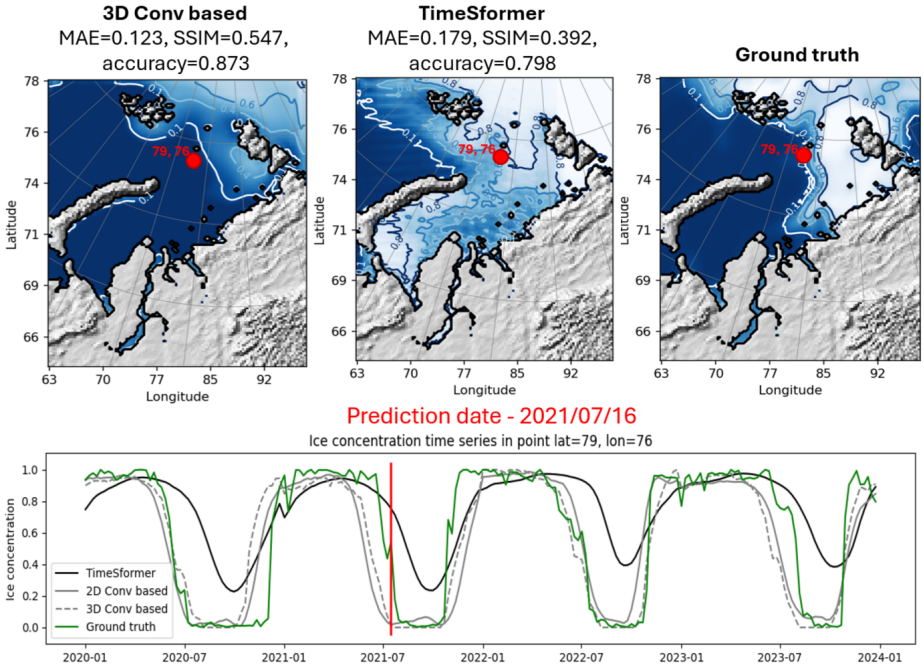

Fig. 4. Comparison of spatial distribution of values on prediction for 2021/07/16 in Kara Sea for different models

Table 2. Time spent on training models on test areas

Model architecture		**2D CNN model**	**3D CNN model**	**SwinLSTM**	**TimeSformer**
Sea	**Image size**	**Train runtime in hours (1000 epochs)***		**Train runtime in hours (120 epochs)***	
Kara	70x60	1.2	2.0	26.2	105.4
Barents	80x75	2.4	3.0	28.7	105.6
East-Siberian	50x62	0.8	1.5	24.6	105.3
Laptev	55x65	1.1	1.7	25.4	105.3
Chukchi	42x73	0.9	1.4	22.1	105.1

* 2D, 3D Conv-based trained on NVIDIA GeForce RTX 4080, SwinLSTM and TimeSformer on NVIDIA Tesla P100 GPU

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

Table 3. Comparison of averaged metrics with SOTA-solutions (SEAS5 and IceNet), 7 month ahead forecast (bold are the best)

Metric	Mean Absolute Error (MAE)				Structural Similarity Index (SSIM)				Accuracy (comparison with ice mask from IceNet)			
Model	**SEAS5**	**2D CNN**	**3D CNN**	**TimeSformer**	**SEAS5**	**2D CNN**	**3D CNN**	**TimeSformer**	**Ice Net**	**2D CNN**	**3D CNN**	**TimeSformer**
Kara Sea	0.093	**0.076**	0.076	0.109	0.653	**0.683**	0.663	0.581	0.918	**0.945**	0.943	0.929
Barents Sea	0.073	0.063	**0.060**	0.129	0.634	**0.684**	0.672	0.489	0.906	0.922	**0.944**	0.916
Laptev Sea	0.101	**0.068**	0.072	0.146	0.703	**0.722**	0.706	0.608	0.967	**0.982**	0.980	0.966
East-Siberian Sea	0.098	0.074	**0.069**	0.177	**0.723**	0.718	0.714	0.685	0.980	**0.990**	0.990	0.988
Chukchi Sea	**0.067**	0.075	0.073	0.147	**0.780**	0.713	0.719	0.588	0.974	0.979	**0.981**	0.962

As can be seen from the tables, both the absolute values of ice concentration and the ice edge position predicted by convolution-based models are of better quality than SOTA. Statistical significance was confirmed using the non-parametric one-sided Mann-Whitney test. IceNet, SEAS5, 2D CNN and TimeSformer all have difference with p-value <0.05. 3D CNN and 2D CNN are not different with p-value 0.91. These models can therefore form a foundation for solutions that go beyond the current state-of-the art.

Toy example on periodic video data. To ensure that the problem of transformers in modelling periodic spatio-temporal data is not specific for analyzed case only, we performed an additional experiment on a 10-frame video (gif animation) based on the manga character "Menhera Shoujo Kurumi-chan" [15]. The animation was divided into frames, scaled to 45x45 resolution, transformed from RGB to 1-channel gray scale with values from 0 to 1. To imitate spatio-temporal data, 10 frames were repeated 5 times, a train set was formed with a slide window on this time series. As a pre-history 20 images were used, the prediction horizon was 10 images of the series ahead. Experiment run with TimeSformer architecture and 2D CNN architecture. The prediction results are shown in Figure 5. Statistical significance of models errors difference confirmed with Mann-Whitney test (p-value for MAE - 0.002, for SSIM - 0.001).

Due to the small training set, we were able to run TimeSformer for 4000 epochs, the CNN was trained for 100000 epochs, the stopping criterion is the number of epochs without L1loss improvement. Detailed convergence plots are described in the supplementary materials (*Media data convergence*).

As the data has an explicit periodicity and no stochastic component, it is expected that the models will be able to approximate the training sample with near-zero error. However, TimeSformer reaches a plateau at 0.02 and produces artifacts in the center of the image. The CNN model converges asymptotically to zero error. Both models capture the temporal dynamics of contour changes well,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

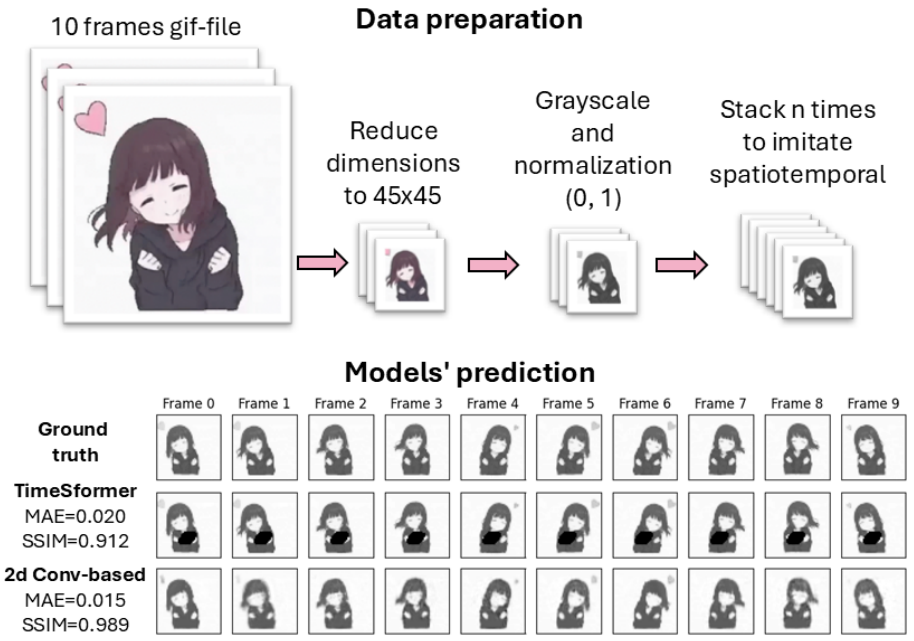

Fig. 5. Media images pre-processing and prediction result (for TimeSformer, CNN-2D)

but the Transformer reproduces the distribution of values within the contour poorly. This behavior is similar to the results obtained with ice data - the model tends to reproduce particularities while losing sight of the more general trend, or vise versa.

6 Conclusion

The results of the experiments confirmed the hypothesis about the limited applicability of transformers for spatio-temporal data with strong periodicity.

For the sea ice concentration forecasting task, the adaptation of the TimeSformer architecture showed a weak reproduction of the time component, due to which the ice concentration in the water area in the summer period did not fall below 0.3, making such a forecast inapplicable. SwinLSTM, aimed at the video prediction task, proved to be helpless in reproducing the annual dynamics. It showed a tendency to self-repeat - for a 7-day forecast a year ahead, summer ice conditions were indistinguishable from the model's initial conditions. It is also worth noting the grainy artifacts in the predictions of transformer-based models.

At the same time, shallow baseline models based on CNN showed reasonable quality compared to SOTA solutions in the field of sea ice prediction. For the forecasts of ice concentration we achieve a quality improvement of up to 30%

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

(Laptev Sea) against SEAS5 system. For ice edge position prediction we achieved comparable results against data-driven system IceNet in terms of accuracy (at the 0.2 threshold) and quality improvements of up to 4-5% in certain water areas (Kara Sea, Barents Sea).

The video prediction experiments for synthetic periodic data also confirm the existence of highlighted problem - convolutional baseline outperforms TimeSformer by 8% for SSIM and 25% for MAE. While further evaluation of the limitations of transformers is still required, we can claim to have provided the solid empirical conformations on the previously poorly discussed problem.

Acknowledgments. The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (project No. № FSER-2024-0004).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andersson, T.R., Hosking, J.S., Pérez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D.C., Wilkinson, J., Phillips, T., et al.: Seasonal arctic sea ice forecasting with probabilistic deep learning. Nature communications **12**(1), 5124 (2021)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
4. Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al.: Aurora: A foundation model of the atmosphere. arXiv preprint arXiv:2405.13063 (2024)
5. Bushuk, M., Msadek, R., Winton, M., Vecchi, G., Yang, X., Rosati, A., Gudgel, R.: Regional arctic sea–ice prediction: Potential versus operational seasonal forecast skill. Climate Dynamics **52**, 2721–2743 (2019)
6. Destro, M., Gygli, M.: Cyclecl: Self-supervised learning for periodic videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2861–2870 (2024)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Grigoryev, T., Verezemskaya, P., Krinitskiy, M., Anikin, N., Gavrikov, A., Trofimov, I., Balabin, N., Shpilman, A., Eremchenko, A., Gulev, S., et al.: Data-driven short-term daily operational sea ice regional forecasting. Remote Sensing **14**(22), 5837 (2022)
9. Guen, V.L., Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11474–11484 (2020)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8

10. Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., et al.: Seas5: the new ecmwf seasonal forecast system. Geoscientific Model Development **12**(3), 1087–1117 (2019)
11. Kim, Y.J., Kim, H.c., Han, D., Stroeve, J., Im, J.: Long-term prediction of arctic sea ice concentrations using deep learning: Effects of surface temperature, radiation, and wind conditions. Remote Sensing of Environment **318**, 114568 (2025)
12. Li, W., Hsu, C.Y., Tedesco, M.: Advancing arctic sea ice remote sensing with ai and deep learning: now and future. EGUsphere **2024**, 1–36 (2024)
13. Liu, Y., Bogaardt, L., Attema, J., Hazeleger, W.: Extended-range arctic sea ice forecast with convolutional long short-term memory networks. Monthly Weather Review **149**(6), 1673–1693 (2021)
14. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
15. Pom: Menhera shoujo kurumi-chan (2018), `https://pom-official.jp/menhera_kurumichan/`
16. Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N.: Weatherbench: a benchmark data set for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems **12**(11), e2020MS002203 (2020)
17. Razzhigaev, A., Mikhalchuk, M., Goncharova, E., Gerasimenko, N., Oseledets, I., Dimitrov, D., Kuznetsov, A.: Your transformer is secretly linear. arXiv preprint arXiv:2405.12250 (2024)
18. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
19. Tan, C., Li, S., Gao, Z., Guan, W., Wang, Z., Liu, Z., Wu, L., Li, S.Z.: Openstl: A comprehensive benchmark of spatio-temporal predictive learning. Advances in Neural Information Processing Systems **36**, 69819–69831 (2023)
20. Tang, S., Li, C., Zhang, P., Tang, R.: Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13470–13479 (2023)
21. Tonboe, R., Lavelle, J., Pfeiffer, R.H., Howe, E.: Product user manual for osi saf global sea ice concentration. Danish Meteorological Institute: Copenhagen, Denmark (2016)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
23. Yarotsky, D.: Error bounds for approximations with deep relu networks. Neural networks **94**, 103–114 (2017)
24. Yu, W., Lu, Y., Easterbrook, S., Fidler, S.: Efficient and information-preserving future frame prediction and beyond. In: International Conference on Learning Representations (2020)
25. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 11121–11128 (2023)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_8