# SIFusion: A Unified Fusion Framework for Multi-granularity Arctic Sea Ice Forecasting

**Jingyi Xu**[1,*], **Shengnan Wang**[1,*], **Weidong Yang**[1,†], **Keyi Liu**[1], **Yeqi Luo**[1], **Ben Fei**[3,†], **Lei Bai**[2]
[1]College of Computer Science and Artificial Intellegence, Fudan University
[2]Shanghai Artificial Intelligence Laboratory, [3]Chinese University of Hong Kong
jyxu22@m.fudan.edu.cn, wdyang@fudan.edu.cn,
benfei@cuhk.edu.hk, baisanshi@gmail.com

## Abstract

Arctic sea ice performs a vital role in global climate and has paramount impacts on both polar ecosystems and coastal communities. In the last few years, multiple deep learning based pan-Arctic sea ice concentration (SIC) forecasting methods have emerged and showcased superior performance over physics-based dynamical models. However, previous methods forecast SIC at a fixed temporal granularity, e.g. sub-seasonal or seasonal, thus only leveraging intra-granularity information and overlooking the plentiful inter-granularity correlations. Specifically, inter-granularity correlations mean that SIC at various temporal granularities exhibits cumulative effects and are naturally consistent, with short-term fluctuations potentially impacting long-term trends and long-term trends provide effective hints for facilitating short-term forecasts in Arctic sea ice. Therefore, in this study, we propose to cultivate temporal multi-granularity that naturally derived from Arctic sea ice reanalysis data and provide a unified perspective for modeling SIC via our **S**ea **I**ce **Fusion** framework. SIFusion is delicately designed to leverage intra-granularity and inter-granularity information to capture granularity-consistent representations that promote forecasting skills. Our extensive experiments indicate that SIFusion outperforms off-the-shelf fixed temporal granularity SIC forecasting deep learning models for their specific temporal granularity.

## 1 Introduction

Arctic sea ice has a profound influence on both local and global climate systems. The near-surface air temperature of Arctic regions has increased at a speed that is two to nearly four times faster than the global average from 1979 to 2021, a phenomenon known as "Arctic amplification" [1, 2]. This accelerated temperature rise performs a key role in the unprecedented rapid diminishing of Arctic sea ice which has extensive consequences that could transcend the polar area. For example, the accelerated reduction of Arctic sea ice could not only jeopardize the survival of species residing in polar regions but also pose adverse effects on local communities whose livelihoods and well-being depend on those animals; it could substantially affect mid-latitude summer weather by weakening the storm tracks [3]; and it will bring new opportunities for marine transportation and new access to natural resources like fossil fuels [4].

Due to the vital role of Arctic sea ice in coastal communities, global climate, and potential impacts on the world's economy, numerical and statistical models have been proposed to forecast pan-Arctic sea ice concentration (SIC) ranging from sub-seasonal to seasonal scale [5, 6]. However, numerical

*Equal Contributions.
†Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

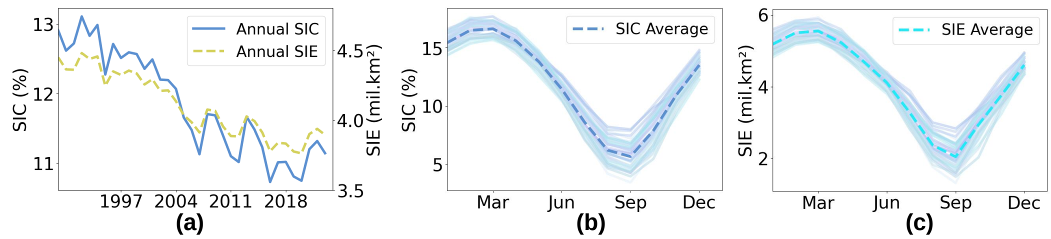


Figure 1: **Visualization of Arctic sea ice trends.** (a)The annual average SIC and SIE trend over the last 35 years (1987-2023); the monthly cyclic trend of SIC **(b)** and SIE **(c)**. Note that the averaged SIC values are calculated over the entire pan-Arctic region, which could only be used to observe the trend.

and statistical models usually rely on high-performance computing on CPU clusters and often lead to complex debugging processes and uncertain parameterization, which limits their performance in forecasting long-term SIC changes. With the advent of deep learning models and their powerful capability in capturing complex patterns within high dimensional data, recent studies have developed end-to-end SIC forecasting models based on deep learning approaches and have presented a promising performance that exceeds previous numerical and statistical methods [7, 8]. Considering existing deep learning-based methods mainly focus on predicting SIC at a specific temporal granularity, e.g., 7 days or 6 months' averages, the intra-granularity correlations are well captured while the inter-granularity information that contains intrinsic annual cyclic trend and intra-seasonal predictability of Arctic sea ice [9] are overlooked.

Over the last few decades, the Arctic sea ice extent (SIE, where SIC value is larger than 15%) has exhibited a continuous declining trend and a clear recurrent variational pattern. For example, the annual pan-Arctic sea ice edge usually starts to expand after the summer melting season in September (Figure 1(b)). Given these patterns, concurrently utilizing inter-granularity and intra-granularity information and employing a unified fusion framework could be mutually beneficial for modeling each granularity. For instance, long-term trends in weekly granularity could be helpful in calibrating short-term daily predictions, and finer granularity features could provide more accurate initial conditions to facilitate seasonal forecasting. Besides, the essence of predicting future SIC is to forecast spatially correlated time series. Their sequentially varying nature requires effective modeling of SIC sequences. However, the most commonly utilized U-Net architecture [10] in previous work [7] implicitly fulfills sequential modeling by channel-wise fusion operations which could be ill-posed for sequence modeling for two reasons: (1) The expansion and contraction of channels in the up-sampling and down-sampling steps disturb the intrinsic sequential feature and complicates the capturing of sequential correlations. (2) When jointly modeling with climate variables, for instance sea surface temperature, fusing different variable channels all together could further corrupt the modeling of spatially correlated time series. Alternatively, adopting explicit modeling of SIC sequences based on sufficiently extracted spatial features could facilitate the spatio-temporal forecasting task.

Based on the above-mentioned motivations, we propose a unified fusion framework for multi-granularity Arctic **S**ea **I**ce **Fusion** forecasting based on Transformer backbone (**SIFusion**). Unlike previous approaches (as demonstrated in Figure 2), we propose to independently tokenize spatial features, explicitly extract sequential information and jointly model three granularities: daily, weekly average, and monthly average. Specifically, SIFusion first embeds SIC from each temporal granularity into independent spatial tokens and sequentially concatenates them to represent temporal fluctuations within each granularity. Then, we treat these independent sequences as correlated granularity variates and utilize the attention mechanism in conjunction with the feed-forward network (FFN) for extracting both intra-granularity and inter-granularity correlations. By incorporating multi-granularity representation, SIFusion could simultaneously generate future SIC in three different temporal scales, boosting not only the performance in a specific temporal scale but also the overall forecasting skill. Our contributions are three-fold:

- We revisit the potentially overlooked inter-granularity information by previous methods for Arctic SIC forecasting and explore the effectiveness of independent spatial tokens representation of SIC for facilitating accurate predictions.
- We propose SIFusion that leverages independent spatial tokenization of SIC and effectively unifies three temporal granularities that cover from sub-seasonal to seasonal scale for better overall representation and improved forecasting performance.

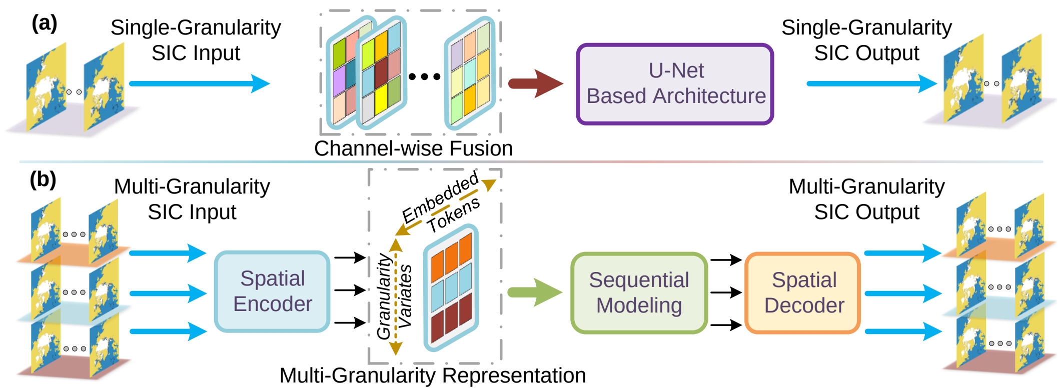


Figure 2: **The main differences** between **(a)** existing mainstream SIC forecasting approaches and **(b)** our SIFusion are follows: (1) Previous models take a channel-wise fusion to jointly extract spatial features, e.g., utilizing 2D convolution to expand and downsample SIC channels. In our case, we focus on capturing an effective spatial tokens representation of SIC by the shared spatial encoder. (2) The correlation among input SIC sequence is implicitly modeled via the U-Net-based architecture in **(a)** while SIFusion explicitly captures intra-granularity and inter-granularity correlation via sequential modeling. (3) We propose leveraging multi-granularity information that is naturally derived from the SIC and embedding it into granularity variates to improve overall forecasting skills.

- The comprehensive experiments demonstrate that by adopting the approach of multi-granularity fusion, our SIFusion achieves state-of-the-art on prediction in each granularity, which advances toward a more practical Arctic sea ice forecasting system.

## 2 Related Works

**Sea ice concentration forecasting.** Researchers have proposed various approaches to forecasting SIC, encompassing numerical and statistical models [11, 12]. However, numerical and statistical models usually rely on the high-performance computing of the CPU cluster and tend to result in complex debugging processes and uncertain parameterization. Recently, deep learning models have drawn the attention of sea ice research communities and have been widely investigated for Arctic sea ice forecasting [13, 14, 15, 16]. These methods utilize U-Net-based architectures to solve daily (SICNet [8], or monthly (IceNet [7], MT-IceNet [16]) SIC forecasting. However, although these U-Net-based architectures are built on top of LSTM [17] or CNN [7], the temporal information inherent in sea ice modeling can not be fully exploited. Moreover, these methods and the latest Transformer-based model [18] concentrate on single-granularity SIC forecasting, where the inter-granularity information from multi-granularity sea ice modeling is overlooked.

**Multi-scale representative learning.** The multi-scale phenomenon is common in vision tasks, while it is always overlooked in sea ice modeling. To exploit the information in multi-scale sources, multi-scale features are commonly exploited by using spatial pyramids [19], dense sampling of windows [20], and the combination of them [21] in the vision community. The learning of CNN-based multi-scale representations is typically approached in two ways: utilizing external factors like multi-scale kernel architectures and multi-scale input architectures [22], or designing internal network layers with skip and dense connections [23]. Recently, there has been a surge of interest in applying transformer-based architectures to computer vision tasks, with the Vision Transformer (ViT) being particularly successful in balancing global and local features compared to CNNs [24]. When revisiting the task of forecasting sea ice concentration, its multiscale features stem from different temporal resolutions. Existing methods focus on a single scale, such as daily, weekly, or monthly. However, different temporal resolutions are inherently connected, and treating them as a single scale for modeling would increase the complexity of network learning.

## 3 SIFusion for Multi-granularity Arctic Sea Ice Forecasting

Given historical Arctic SIC records $Y = \{X_{T-L-1}, ..., X_{T-1}, X_T\} \in [0\%, 100\%]^{L\times H\times W}$, where $L$ is the input length of a specific granularity which includes the given observation time step $T$, $H$

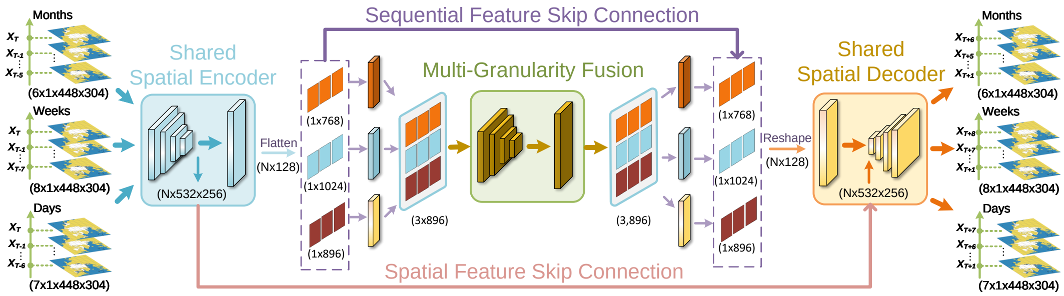


Figure 3: **Overview of proposed SIFusion**, which comprises three main components: **(1)** The **shared spatial encoder** first independently extracts spatial features of input SIC from each granularity (i.e. 7 days, 8 weeks' averages and 6 months' averages) to obtain spatial tokens, and then concatenates these spatial tokens accordingly. **(2)** The embedded spatial tokens are subsequently flattened with respect to their granularity and linearly projected into the same length. We propose to utilize an encoder-only transformer backbone to perform **multi-granularity fusion** which explicitly captures both inter-granularity and intra-granularity sequential features. **(3)** Lastly, the predicted multi-granularity features are restored to the shape of the input via linear transformation and the **shared spatial decoder**.

and $W$ denotes the rectangle pan-Arctic region, the forecasting model predicts the subsequent SIC values $\hat{Y} = \{X_{T+1}, ..., X_{T+P-1}, X_{T+P}\} \in [0\%, 100\%]^{P \times H \times W}$ with forecasting lead times of $P$. In this study, our SIFusion jointly models three granularities, i.e., daily, weekly average, and monthly average SIC values that cover both sub-seasonal and seasonal variations, and simultaneously forecasts on all these temporal scales. For each temporal granularity, the input length $L$ equals the forecasting lead times $P$. The overview of the proposed SIFusion architecture is presented Figure 3. The shared spatial encoder and decoder perform SIC tokenization and restoration while multi-granularity fusion explicitly extracts sequential information.

### 3.1 Sea ice concentration tokenization

Existing mainstream deep learning-based methods for SIC forecasting adopt U-Net architectures and leverage 2D convolution to perform channel-wise expansion and downsampling that extracts both spatial features and temporal dependencies. However, since U-Net-based models are not specifically designed for sequence modeling [27], the joint spatial-channel fusion of SIC and implicit sequence modeling could be ill-posed properties for spatio-temporal forecasting tasks. In this regard, we propose to independently tokenize spatial features at first, which could disentangle the above ill-posed problem and be beneficial for SIC forecasting.

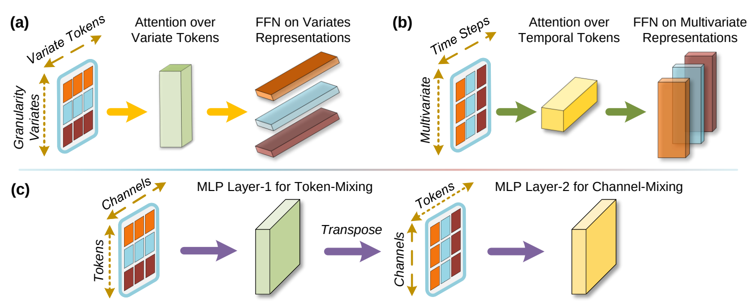


Figure 4: **Comparison between different backbones for temporal sequence modeling**: **(a)** Our proposed SIFusion sequentially concatenates independent SIC tokens that are derived from each temporal scale as a granularity variate and applies an attention mechanism over the embedded variate tokens. The FFN transforms the variate representation for the input of the next layer; **(b)** For vanilla Transformer architecture [25], it applies an attention mechanism over temporal tokens and FFN is applied on multivariate representations; **(c)** The MLP-mixer [26] approach first performs token-wise mixing, then transposes the extracted features to apply channel-wise mixing. The vanilla Transformer and MLP-mixer both fall short of modeling the sequential information of sea ice.

**Independent spatial embedding.** Since we aim to simultaneously model SIC derived from three temporal granularities, encoding their spatial features into shared embedding space not only yields consistent representation but also reduces the number of trainable parameters. Inspired by prior works [28, 29], we utilize Swin Transformer V2 [30] as the backbone for both shared spatial encoder and decoder.

Specifically, each SIC input is independently fed into the shared spatial encoder and partitioned by a non-overlapped window to generate patch representation [24] with 32 spatial channels (the original SIC data has only one channel). To preserve local SIC information, we choose the smallest 2 by 2 window size for the patch partition. Then, the patch tokens are further transformed by the first two Swin Transformer blocks. The multi-scale spatial features are extracted through the subsequent hierarchical encoder layers which comprise a patch merging operation and two Swin Transformer blocks. The patch merging operation first concatenates the spatial feature of each group of 2 by 2 adjacent patch representations from the previous encoder layer. The calculation of each pair of two consecutive Swin Transformer blocks in encoder layers can be described as follows:

$$\begin{aligned}
&z_s^b = LN(WMSA(z^{b-1})) + z^{b-1}, \\
&z^b = LN(MLP(z_s^b)) + z_s^b, \\
&z_s^{b+1} = LN(SWMSA(z^b)) + z^b, \\
&z^{b+1} = LN(MLP(z_s^{b+1})) + z_s^b,
\end{aligned} \tag{1}$$

where $z_s^b$ and $z^b$ represents the output spatial feature of the (**S**hifted) **W**indow-**M**ulti-head **S**elf **A**ttention module and the MLP module for block $b$, respectively; LN denotes the layer normalization operation [31]. The attention mechanism with a shifted window could effectively extract neighboring SIC information and sufficiently represent the local correlation of sea ice. After all input SIC are independently encoded into 2D spatial features, we apply linear projection to generate 1D tokens for each SIC to obtain compact spatial representation for sequential modeling.

The shared spatial decoder adopts an identical Swin Transformer backbone and the decoding procedure is symmetrical to the encoding process, except that the patch merging operation is replaced by the patch expanding operation [32]. While patch merging downsamples the input spatial feature dimension and increases the embedding channels, patch expanding symmetrically restores the resolution of the feature map and merges channels via linear transformation.

**Spatial feature skip connection.** Since the SIC features encoded by Swin Transformer blocks will be tokenized into highly compact sequence representation, the spatial SIC information should be maximally preserved during the sequential modeling. Besides, our proposed sequential modeling backbone comprises deep encoding layers which might lead to loss of embedded spatial features. To preserve spatial SIC information and avoid insufficient restoration, we propose to add a skip connection between the output of the last pair of Swin Transformer blocks in the spatial encoder and the input of the first block in the shared decoder (see in Figure 3).

### 3.2 Multi-granularity fusion

We propose to jointly model three granularities that cover sub-seasonal to seasonal scale, i.e., 7 days, 8 weeks averages, and 6 months averages, and explicitly capture inter-granularity correlation and intra-granularity sequential information.

**Modeling granularity variates.** As mentioned in Section 3.1 the shared spatial encoder transforms each SIC into independent 1D tokens. These individual spatial tokens are then concatenated sequentially based on their granularity respectively and utilized to form the multi-granularity representation. As each granularity incorporates a different time span, the dimensions of concatenated granularity sequences are mismatched. Considering that both the weekly average and monthly average are derived from daily SIC values, we choose to tokenize those sequences further and align their feature dimensions with the length of daily input using a linear transformation as follows:

$$g_f = Linear(vTokenizer(SIC_{granularity})), \tag{2}$$

where $granularity \in$ [daily, weekly, monthly], $g_f$ is the aligned granularity feature, $vTokenizer$ is the shared Swin Transformer spatial encoder.The generated multi-granularity variates are subsequently fed into the sequential modeling backbone. Encouraged by prior work [33], we propose to adopt an encoder-only Transformer architecture as the sequential modeling backbone for multi-granularity fusion in Figure 3 that: (1) applies multi-head self-attention on the embedded granularity variate tokens to explicitly capture inter-granularity correlations; (2) each granularity variate is independently processed by FFN to extract intra-granularity information (as depicted Figure 4(a)). As for the conventional usage of vanilla Transformer in sequence prediction, the attention mechanism is applied on embedded temporal tokens which comprise variate information collected from the same time step (as in Figure 4(b)). The vanilla Transformer is challenged in forecasting series with larger lookback windows due to performance degradation and computation explosion. Furthermore, the temporal token embeddings incorporate multiple variates that represent distinct physical measurements, which may struggle to capture variate-specific representations and potentially lead to the generation of incoherent attention maps. However, in sea ice modeling, each dimension of the tokenized granularity variate incorporates SIC features that come from a different time span. This could lead to poor representation of sequential SIC features and restrict the effective modeling of inter-granularity correlations. Experimentally, we will show in Section 4.2 that by adopting our sequential modeling, the overall performance is superior to alternative backbones. After each SIC granularity variate token is properly fused and encoded, the final prediction of future granularity variate features is generated through a linear projection layer.

**Sequential feature skip connection.** Considering the concatenated sequence of SIC features are linearly transformed and aligned to form the multi-granularity variate representation, the significant original sequential feature needs appropriate preservation. Besides, the deep sequence encoding process could introduce unintended noise that deteriorates the modeling of intra-granularity correlation. To compensate for the intra-granularity information and reduce the potential impact that impairs inter-granularity modeling, we propose to utilize the cross-attention mechanism as a sequential skip connection (in Figure 3), where the latent query features are sourced from the concatenated sequence token before the linear projection and the predicted SIC sequence generates both key and value latent representations. The details about this process can be found in Appendix.

# 4 Experiments

In this section, we evaluate the forecasting performance of our SIFusion over 8 years of SIC data and compare it with other deep learning models. The implementation details and evaluation metrics calculation of our SIFusion are provided in Appendix.

**Datasets.** We evaluate our proposed SIFusion framework on the G02202 Version 4 dataset from the National Snow and Ice Data Center (NSIDC). It records daily SIC data starting from October $25^{th}$ 1978 and provides the coverage of the pan-Arctic region (N:89.8°, S:31.1°, E:180°, W:$-180$°). Each daily SIC data is formed of 448 x 304 pixels and each pixel corresponds to the area of a 25km x 25km grid. The SIC data has a range of 0% to 100% and areas where SIC value is greater than 15% indicate the SIE. We choose data from October $25^{th}$ 1978 to the end of 2013 as the training dataset, the years 2014 and 2015 are selected as the validation set, and data collected from 2016 to 2023 are used to test models.

**Data curation.** The official data was preserved in the NetCDF format. We use the open-source Python package 'netCDF4' to read data from the file with the suffix '.nc'. The sea ice concentration data can then be extracted by applying the official variable name, i.e., 'cdr_seaice_conc', to the API.

**Evaluation metrics.** This study employs root mean square error (RMSE) and mean absolute error (MAE) to assess forecasting accuracy, along with the $R^2$ score to quantify model performance. For SIE prediction, the Integrated Ice-Edge Error score [34] metric is used, which decomposes errors into overestimation (O) and underestimation (U) components. Additionally, the SIE difference (SIE$_{dif}$) calculates the absolute discrepancy between predicted and observed ice area (in millions of $km^2$). The Nash-Sutcliffe Efficiency [35] further evaluates prediction quality by comparing model performance to a baseline mean. The detailed calculations could be found in the Appendix A.1.

## 4.1 Multi-granularity forecasting

**Baselines.** Since our SIFusion simultaneously generates predictions of three granularities, we select corresponding forecasting deep learning-based models for comparison. Specifically, we re-implemented SICNet [8] and trained under an identical environment for direct comparison on 7 days SIC forecasting; Due to dataset and code accessibility, we adopt performance of sub-seasonal forecasting methods as SICNet$_{90}$ [36], IceFormer [18], and seasonal forecasting methods IceNet [7],

Table 1: **Quantitative results of SIC forecasting**. We compare the performance of SIFusion in each temporal granularity with the corresponding deep learning based methods.

| Temporal Scale | Lead Times | Methods | RMSE↓ | MAE↓ | $R^2$↑ | NSE↑ | IIEE↓ | $\text{SIE}_{dif}$↓ |
|---|---|---|---|---|---|---|---|---|
| Sub-seasonal | 7 Days (Daily) | SICNet [8] | 0.0490 | 0.0100 | 0.982 | 0.979 | 976 | 0.0718 |
| | | ConvLSTM [37] | 0.0681 | 0.0263 | 0.971 | - | - | - |
| | | PredRNN [38] | 0.0594 | 0.0220 | 0.977 | - | - | - |
| | | SimVP [39] | 0.0640 | 0.0238 | 0.974 | - | - | - |
| | | **SIFusion** | **0.0429** | **0.0096** | **0.987** | **0.985** | **926** | **0.0380** |
| | 45 Days (Daily) | IceFormer [18] | 0.0660 | 0.0201 | 0.960 | - | - | - |
| | 90 Days (Daily) | $\text{SICNet}_{90}$ [36] | - | 0.0512 | - | - | - | - |
| | 8 Weeks Average (Weekly) | **SIFusion** | **0.0625** | **0.0140** | **0.973** | **0.968** | **1600** | **0.1541** |
| Seasonal | 6 Months Average (Monthly) | IceNet [7] | 0.1820 | 0.0916 | 0.567 | - | - | - |
| | | MT-IceNet [16] | 0.0777 | 0.0197 | 0.915 | - | - | - |
| | | **SIFusion** | **0.0692** | **0.0166** | **0.917** | **0.910** | **2156** | **0.2083** |

MT-IceNet [16] that reported in the original paper for reference. We also include ConvLSTM [37], PredRNN [38] and SimVP [39] for comparison.

**Main results.** The overall performance of SIFusion and baseline methods is listed in 1. The lower RMSE/MAE indicates a more accurate forecast in SIC values. Methods with lower IIEE/$\text{SIE}_{dif}$ are more capable of identifying the edge of sea ice while higher $R^2$/NSE suggests that the predicted spatial patterns are closer to the truth of the ground. Our proposed method achieves the best performance in all metrics for forecasting 7 days SIC, establishes a new state-of-the-art method for sub-seasonal weekly average prediction, and presents superior seasonal SIC forecasting capability. Considering the fact that baseline methods, except for SICNet, utilize several additional atmospheric and oceanic variables to facilitate forecasting, and our SIFusion only leverages SIC data with carefully extracted intrinsic inter-granularity correlation, it verifies the effectiveness of the proposed approach for multi-granularity forecasting.

**Qualitative analysis.** To visually verify the forecasting skills of SIFusion, we select the end of the melting season in September 2022. From Figure 1(a) we can observe that the annual Arctic sea ice in 2022 has increased by a considerable margin, which is against the persisting long-term declining trend. This unusual rise makes SIC and SIE difficult for our model to predict since it only learns from the data collected before 2014. Starting from September $1^{st}$, we calculate the averaged SIC of 7 days, 4 weeks and 1 month that correspond to three temporal granularities of SIFusion. The ground truth of calculated average SIC along with the ground truth and predicted SIE are visualized in Figure 5. The forecasting results in the lower row are produced by SIFusion and the upper row represents predictions generated by three variants of SIFusion that only leverage single-granularity SIC, we will discuss later in Section 4.2.

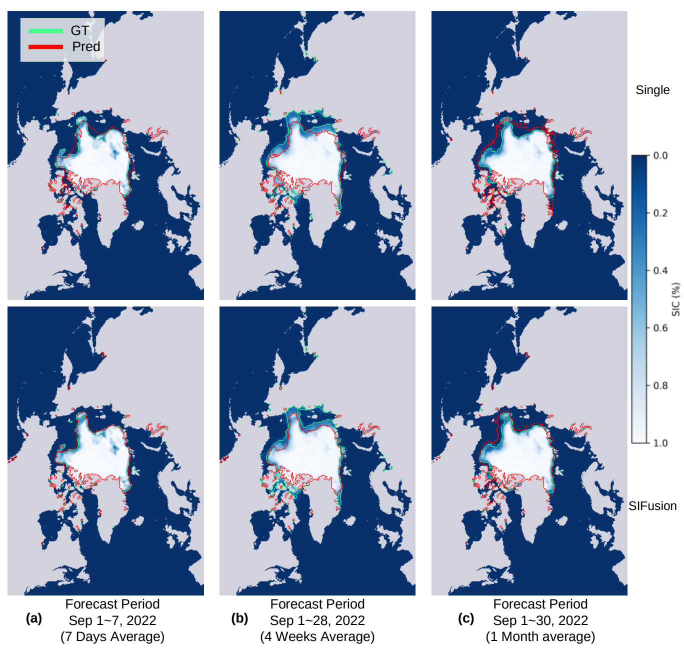


Figure 5: **Qualitative analysis of SIE prediction.** The derived SIE ground truth and prediction generated by SIFusion and three single-granularity models (one for each temporal granularity) over: **(a)** The first week of September; **(b)** 4 weeks; **(c)** 1 month. Considering the abnormal increase of Arctic sea ice in 2022, our proposed method could still produce reasonable forecasts that keep the similar overall shape of Arctic SIE.

Despite the inconsistent annual trend of Arctic SIC in 2022, our

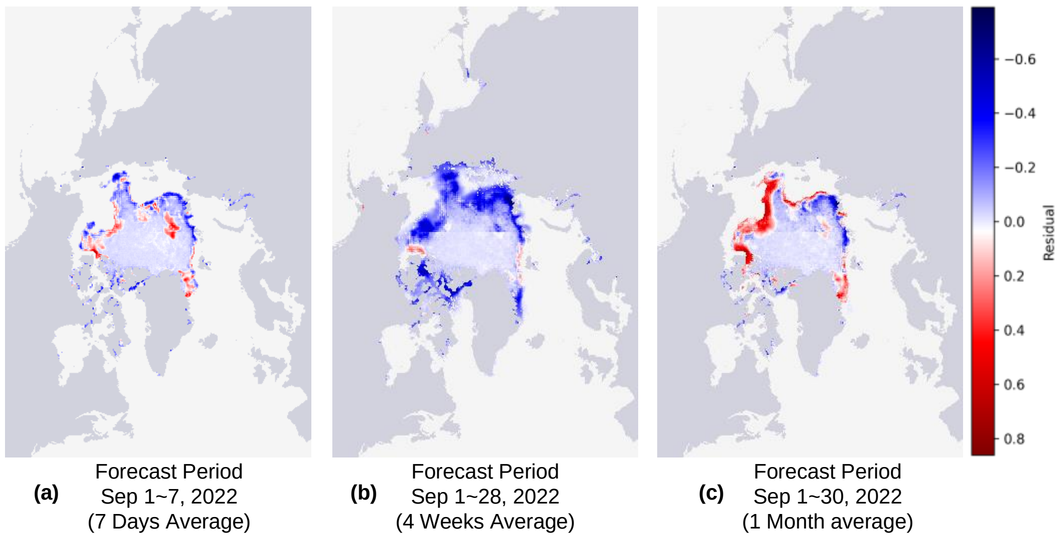


Figure 6: **Spatial residual of predicted SIC.** We examine the spatial patterns of forecasting results over the same period presented in Figure 5: SIFusion could generate consistent daily forecasts **(a)**. Considering the abnormal Arctic SIC change in 2022, the annual trend could be different than the SIC data on which the model was trained, SIFusion could still predict weekly **(b)** and monthly **(c)** average SIC with a bounded residual region rather than scattered forecasting results. The spatial residual is calculated by using predicted SIC to subtract the ground truth value.

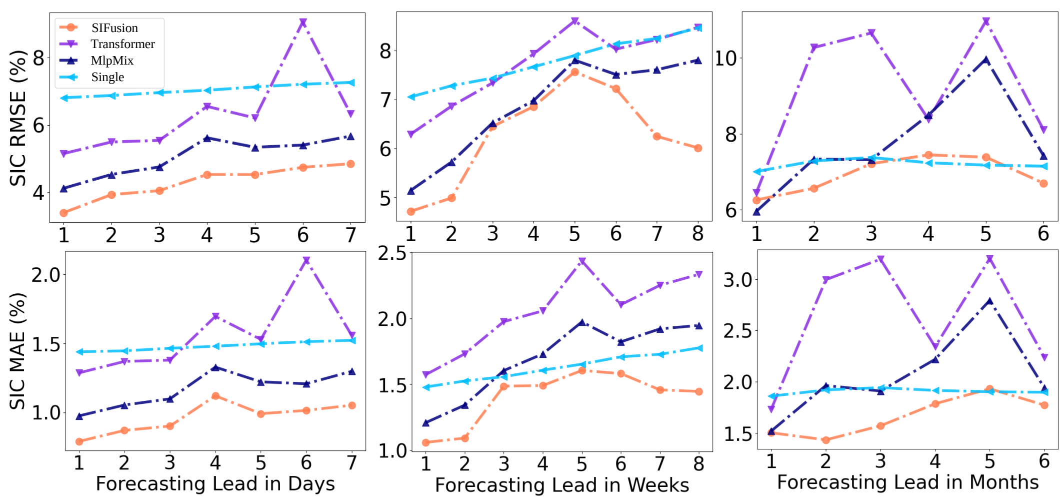


Figure 7: **Averaged intra-granularity forecasting error.** We evaluate models trained on multi-granularity and single-granularity SIC and plot RMSE and MAE of each lead time step in three temporal granularities over the test dataset.

method could still generate forecasts that are consistent with the average SIE in the first week of September (Figure 5(a)), and the general shape in both 4 weeks' average (Figure 5(b)) and 1-month average (Figure 5(c)). Compared to models with a similar backbone of SIFusion but only leveraging single-granularity SIC, the prediction of SIE is noticeably different from the ground truth, indicating that SIFusion could effectively leverage multi-granularity SIC to improve forecasting skills.

We plot spatial residuals to further investigate the learned spatial patterns of our SIFusion. In Figure 6(a), SIFusion could accurately predict the first week of SIC, while in coarser weekly average granularity our SIFusion tends to slightly underestimate in Arctic sea ice edge areas (Figure 6(b)). For the predicted monthly average of September 2022, the overall shape of SIE resembles the observation but overestimates SIC along the boundary.

Table 2: **Effectiveness of multi-granularity representation**. *Multi* represents the proposed SIFusion and *Single* stands for models with similar backbone but trained solely on single-granularity data.

| Temporal Scale | Lead Time | Granularity | RMSE↓ | MAE↓ | $R^2$↑ | NSE↑ | IIEE↓ | $\text{SIE}_{dif}$↓ |
|---|---|---|---|---|---|---|---|---|
| Sub-seasonal | 7 Days | Single | 0.0704 | 0.0148 | 0.982 | 0.979 | 1018 | 0.0509 |
| | | Multi | **0.0429** | **0.0096** | **0.987** | **0.985** | **926** | **0.0380** |
| | 8 Weeks Average | Single | 0.0771 | 0.0163 | 0.962 | 0.954 | 2208 | 0.3301 |
| | | Multi | **0.0625** | **0.0140** | **0.973** | **0.968** | **1600** | **0.1541** |
| Seasonal | 6 Months Average | Single | 0.0721 | 0.0191 | 0.882 | 0.873 | 2482 | 0.4298 |
| | | Multi | **0.0692** | **0.0166** | **0.917** | **0.910** | **2156** | **0.2083** |

Table 3: **Effectiveness of proposed approach for multi-granularity fusion**. We adopt conventional utilization of the Transformer and the recent trend in leveraging a full MLP-based backbone [26] for temporal sequence modeling as counterparts of our proposed sequential backbone.

| Temporal Scale | Lead Time | Method | RMSE↓ | MAE↓ | $R^2$↑ | NSE↑ | IIEE↓ | $\text{SIE}_{dif}$↓ |
|---|---|---|---|---|---|---|---|---|
| Sub-seasonal | 7 Days | MLP Mixing | 0.0506 | 0.0117 | 0.984 | 0.981 | 1153 | 0.1265 |
| | | Transformer | 0.0633 | 0.0159 | 0.970 | 0.965 | 1519 | 0.2338 |
| | | SIFusion | **0.0429** | **0.0096** | **0.987** | **0.985** | **926** | **0.0380** |
| | 8 Weeks Average | MLP Mixing | 0.0689 | 0.0169 | 0.969 | 0.963 | 2222 | 0.3839 |
| | | Transformer | 0.0771 | 0.0206 | 0.970 | 0.964 | 1718 | 0.2028 |
| | | SIFusion | **0.0625** | **0.0140** | **0.973** | **0.968** | **1600** | **0.1541** |
| Seasonal | 6 Months Average | MLP Mixing | 0.0775 | 0.0206 | 0.857 | 0.845 | 2477 | 0.3837 |
| | | Transformer | 0.0913 | 0.262 | 0.833 | 0.821 | 3490 | 0.4902 |
| | | SIFusion | **0.0692** | **0.0166** | **0.917** | **0.910** | **2156** | **0.2083** |

### 4.2 Ablation study

To further analyze the performance of our proposed method, we trained five additional variants of SIFusion (as in Figure 7), i.e., three single-granularity models that respectively utilize temporal granularities in SIFusion, and two multi-granularity forecasting models with different backbones to perform the multi-granularity fusion.

**Effectiveness of multi-granularity modeling.** We first verify our proposed multi-granularity modeling approach by comparing SIFusion with models that comprise a similar model architecture but only adopt single granularity SIC data. Comprehensive experiments in Table 2 show that by leveraging the naturally derived multi-granularity SIC, the overall performance in all temporal scales can be promoted by a significant margin. For each individual forecasting lead time, SIFusion consistently outperforms models solely trained on single-granularity data (as shown in Figure 7).

**Alternative backbone for multi-granularity fusion.** To validate the effectiveness of our proposed multi-granularities fusion and sequential modeling approach, we compare the performance of our SIFusion with two other variants that are trained on the identical multi-granularity data with different sequential backbones, i.e., Transformer and MLP mixer [26, 40]. Considering intra-granularity performance in Figure 7, SIFusion presents superior forecasting skill in each time step of daily, weekly average and monthly average when compared to multi-granularity variants, indicating the effectiveness of multi-granularity SIC variates for sequential modeling. As shown in Table 3, our SIFusion outperforms these variants by a great margin, demonstrating the intra-granularity and inter-granularity correlations inherent in the sea ice modeling benefits for the forecasting.

**Adaptation to missing observations and change in temporal scale.** Considering our multi-granularity fusion framework is generic for different temporal scales, it is essential to further explore (1) whether SIFusion applies to missing SIC observations and (2) if it still works when the input temporal scale varies. We first randomly masked the daily scale, i.e., select one of the 7 days and set the input SIC values to a mask token, and then extended the daily scale from 7 days to 10 days. The evaluation of these two variants is presented in Table 4. For the scenario with missing values and changed input daily scale, the performance of all three temporal scales slightly drops, but still outperforms the single-granularity baseline in terms of RMSE and MAE. This could indicate that our SIFusion is applicable to missing observations. By comparing the performance variance between SIFusion and $\text{SIFusion}_{10-Days}$, we could find that the daily and monthly figures are close, but the

Table 4: **Application to missing daily observation and different temporal scale**. To further explore the capability of our multi-granularity fusion framework, we randomly masked daily training data (SIFusion$_{Mask}$) and extended the daily input from 7 days to 10 days (SIFusion$_{10-Days}$). Metrics for SIFusion$_{10-Days}$ are calculated using the first 7-day prediction.

| Temporal Scale | Lead Time | Method | RMSE↓ | MAE↓ | $R^2$↑ | NSE↑ | IIEE↓ | SIE$_{dif}$↓ |
|---|---|---|---|---|---|---|---|---|
| Sub-seasonal | 7 Days | SIFusion$_{Mask}$ | 0.0633 | 0.0141 | 0.975 | 0.971 | 1280 | 0.0941 |
| | | SIFusion$_{10-Days}$ | 0.0643 | 0.0140 | 0.977 | 0.974 | 1292 | 0.0978 |
| | | SIFusion | **0.0429** | **0.0096** | **0.987** | **0.985** | **926** | **0.0380** |
| | 8 Weeks Average | SIFusion$_{Mask}$ | 0.0639 | 0.0162 | 0.969 | 0.963 | 2214 | 0.2674 |
| | | SIFusion$_{10-Days}$ | 0.0664 | 0.0155 | 0.958 | 0.950 | 2335 | 0.3472 |
| | | SIFusion | **0.0625** | **0.0140** | **0.973** | **0.968** | **1600** | **0.1541** |
| Seasonal | 6 Months Average | SIFusion$_{Mask}$ | 0.0719 | 0.0166 | 0.871 | 0.862 | 2258 | 0.3075 |
| | | SIFusion$_{10-Days}$ | **0.0663** | **0.0153** | 0.887 | 0.879 | 2169 | 0.2950 |
| | | SIFusion | 0.0692 | 0.0166 | **0.917** | **0.910** | **2156** | **0.2083** |

identification of sea ice extent on the weekly scale suffers a larger performance drop. Since we use 7 days on a daily scale, i.e., the 7-day time naturally aligns with one week, the results could indicate that keeping this alignment between the daily scale and a one-week token is beneficial. As to the 8 weeks' average at the weekly scale, the time span of 8 weeks is still within a month, so that one one-month token could represent the information of the weekly scale.

## 5 Conclusion and Future Work

In this paper, we propose SIFusion, a transformer-based sea ice concentration forecasting framework that unifies multi-granularity covering from sub-seasonal to seasonal scale to enhance the forecasting skills. Specifically, we propose to explore the independent spatial tokens representation of SIC and explicitly modeling intra-granularity information. These spatial tokens are sequentially concatenated within their own granularity and go through multi-granularity fusion to effectively capture the inter-granularity correlations. Experiments demonstrate that our SIFusion achieves skillful forecasting in each granularity and outperforms methods trained on single-granularity SIC and climate data. **Limitation.** Despite the effectiveness of only using SIC as the input, the absence of climate variables could still limit the performance of the proposed approach for modeling Arctic sea ice. The weekly average granularity at sub-seasonal scale was not fully aligned with existing models, but it provides critical correlations for bridging the gap between short-term and long-term forecasting, our SIFusion sets up a new baseline for future explorations. Since our proposed framework is versatile, the climate information could be easily incorporated in future work. The joint modeling of climate variables and SIC using SIFusion from a multi-granularity perspective could provide a more comprehensive understanding of climate change both within and beyond the Arctic region. Besides, leveraging the hidden knowledge of atmospheric and oceanic dynamics embedded in pre-trained weather and climate foundation models could enhance the forecasting skill and boost the overall performance. Lastly, the capability of SIFusion to simultaneously forecast Arctic SIC at multiple granularities could make it a promising candidate to fulfill the challenging and critical sub-seasonal to seasonal (S2S) sea ice forecast.

## Acknowledgments

This research was supported by Shanghai's Key Technology Development Program, under the project 'Intelligent Integrated Arctic Ice-Snow-Atmosphere Observation Platform and Short- to Medium-Term High-Precision Sea Ice Forecasting' (Grant No. 25DZ3102600), Shanghai Artificial Intelligence Laboratory, and the JC STEM Lab of AI for Science and Engineering, funded by The Hong Kong Jockey Club Charities Trust, the Research Grants Council of Hong Kong (Project No. CUHK14213224).

We want to express our sincere gratitude to Professor Xiaojun Yuan from the Lamont-Doherty Earth Observatory, Columbia University, for her insightful and invaluable help and discussion. We also extend our gratitude to Professor Lei Wang from the Institute of Atmospheric Sciences, Fudan University, whose advice has been instrumental in our study.

## References


[1] James A Screen and Ian Simmonds. The central role of diminishing sea ice in recent arctic temperature amplification. *Nature*, 464(7293):1334–1337, 2010.

[2] Mika Rantanen, Alexey Yu Karpechko, Antti Lipponen, Kalle Nordling, Otto Hyvärinen, Kimmo Ruosteenoja, Timo Vihma, and Ari Laaksonen. The arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment*, 3(1):168, 2022.

[3] Stephen J Vavrus. The influence of arctic amplification on mid-latitude weather and climate. *Current Climate Change Reports*, 4:238–249, 2018.

[4] Warwick F Vincent. Arctic climate change: Local impacts, global consequences, and policy implications. *The Palgrave handbook of Arctic policy and politics*, pages 507–526, 2020.

[5] Stephanie J Johnson, Timothy N Stockdale, Laura Ferranti, Magdalena A Balmaseda, Franco Molteni, Linus Magnusson, Steffen Tietsche, Damien Decremer, Antje Weisheimer, Gianpaolo Balsamo, et al. Seas5: the new ecmwf seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117, 2019.

[6] Lei Wang, Xiaojun Yuan, and Cuihua Li. Subseasonal forecast of arctic sea ice concentration via statistical approaches. *Climate Dynamics*, 52:4953–4971, 2019.

[7] 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, 2021.

[8] Y Ren, X Li, and W Zhang. A data-driven deep learning model for weekly sea ice concentration prediction of the pan-arctic during the melting season, ieee t. geosci. remote, 60, 4304819, 2022.

[9] Lei Wang, Xiaojun Yuan, Mingfang Ting, and Cuihua Li. Predicting summer arctic sea ice concentration intraseasonal variability using a vector autoregressive model. *Journal of Climate*, 29(4):1529–1543, 2016.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[11] Wanqiu Wang, Mingyue Chen, and Arun Kumar. Seasonal prediction of arctic sea ice extent from a coupled dynamical forecast system. *Monthly Weather Review*, 141(4):1375–1394, 2013.

[12] Xiaojun Yuan, Dake Chen, Cuihua Li, Lei Wang, and Wanqiu Wang. Arctic sea ice seasonal prediction by a linear markov model. *Journal of Climate*, 29(22):8151–8173, 2016.

[13] Zisis I Petrou and Yingli Tian. Prediction of sea ice motion with convolutional long short-term memory networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6865–6876, 2019.

[14] Young Jun Kim, Hyun-Cheol Kim, Daehyeon Han, Sanggyun Lee, and Jungho Im. Prediction of monthly arctic sea ice concentrations using satellite and reanalysis data based on convolutional neural networks. *The Cryosphere*, 14(3):1083–1104, 2020.

[15] Sahara Ali, Yiyi Huang, Xin Huang, and Jianwu Wang. Sea ice forecasting using attention-based ensemble lstm. *arXiv preprint arXiv:2108.00853*, 2021.

[16] Sahara Ali and Jianwu Wang. Mt-icenet-a spatial and multi-temporal deep learning model for arctic sea ice forecasting. In *2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, pages 1–10. IEEE, 2022.

[17] Yang Liu, Laurens Bogaardt, Jisk Attema, and Wilco Hazeleger. Extended-range arctic sea ice forecast with convolutional long short-term memory networks. *Monthly Weather Review*, 149(6):1673–1693, 2021.

[18] Qingyu Zheng, Ru Wang, Guijun Han, Wei Li, Xuan Wang, Qi Shao, Xiaobo Wu, Lige Cao, Gongfu Zhou, and Song Hu. A spatio-temporal multiscale deep learning model for subseasonal prediction of arctic sea ice. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. IEEE, 2006.

[20] Shengye Yan, Xinxing Xu, Dong Xu, Stephen Lin, and Xuelong Li. Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 473–487. Springer, 2012.

[21] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[22] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748, 2015.

[23] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[24] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[25] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[26] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

[27] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[28] Yuan Hu, Lei Chen, Zhibin Wang, and Hao Li. Swinvrnn: A data-driven ensemble forecasting model via learned distribution perturbation. *Journal of Advances in Modeling Earth Systems*, 15(2):e2022MS003211, 2023.

[29] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

[30] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022.

[31] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.

[32] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.

[33] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[34] Helge F Goessling, Steffen Tietsche, Jonathan J Day, Ed Hawkins, and Thomas Jung. Predictability of the arctic sea ice edge. *Geophysical Research Letters*, 43(4):1642–1650, 2016.

[35] J.E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970.

[36] Yibin Ren and Xiaofeng Li. Predicting the daily sea ice concentration on a subseasonal scale of the pan-arctic during the melting season by a deep learning model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[37] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[38] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017.

[39] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3170–3180, 2022.

[40] Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 459–469, 2023.

# A Appendix

## A.1 Evaluation Metrics

To evaluate SIFusion, we select widely used root mean square error (RMSE) and mean absolute error (MAE) for comparison of forecasting accuracy. We also leverage $R^2$ score to evaluate the performance:

$$R^2 = 1 - \frac{RSS}{TSS}. \tag{3}$$

where RSS represents the sum of squares of residuals and TSS denotes the total sum of squares. The Integrated Ice-Edge Error score [34] is introduced to evaluate the prediction of SIE:

$$IIEE = O + U, \tag{4}$$

$$O = \sum(Max(SIE_p - SIE_t, 0)), \tag{5}$$

$$U = \sum(Max(SIE_t - SIE_p, 0)), \tag{6}$$

$$SIE_p, SIE_t = \begin{cases} 1, SIC > 15 \\ 0, SIC \leq 15 \end{cases} \tag{7}$$

where O and U represent the overestimated and underestimated SIE between the prediction ($SIE_p$) and the ground truth ($SIE_t$), respectively. The difference between the forecasted and ground truth sea ice area (in millions of $km^2$) is calculated as follows:

$$SIE_{dif} = \frac{\sum(|SIE_p - SIE_t|) \times 25 \times 25}{1000000}. \tag{8}$$

We also adopt the Nash-Sutcliffe Efficiency [35] to further evaluate the predicted quality:

$$NSE = \frac{1 - \sum((SIC_t - SIC_p)^2)}{\sum((SIC_t - Mean(SIC_t))^2)} \tag{9}$$

## A.2 The details of sequential feature skip connection

$$\begin{gathered} CrossAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V, \\ Q = W_Q^g \cdot z_{in}^g, K = W_K^g \cdot z_{pred}^g, V = W_V^g \cdot z_{pred}^g \end{gathered} \tag{10}$$

where $g$ denotes each granularity. $z_{in}^g, z_{pred}^g \in \mathbb{R}^{1\times d_z}$ represents the sequential features before linear projection and the prediction, respectively. $W_Q^g, W_K^g, W_V^g \in \mathbb{R}^{d\times d_z}$ are the query, key and value projection matrices.

## A.3 Visualization of forecasting results

In this section, we will present more visualization of forecasting results generated by SIFusion.

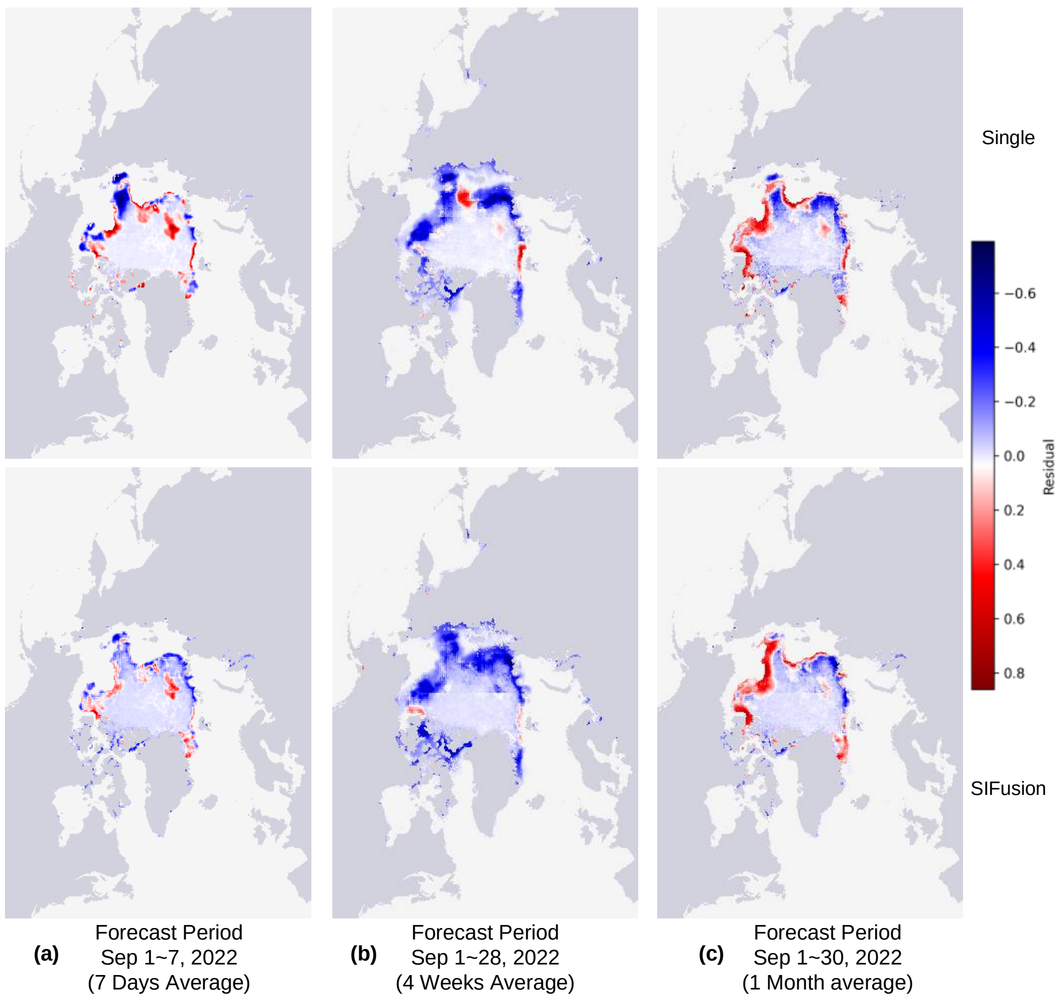


Figure 8: **Spatial residual comparison.** We compare the spatial patterns of forecasting results produced by SIFusion and single-granularity variants.

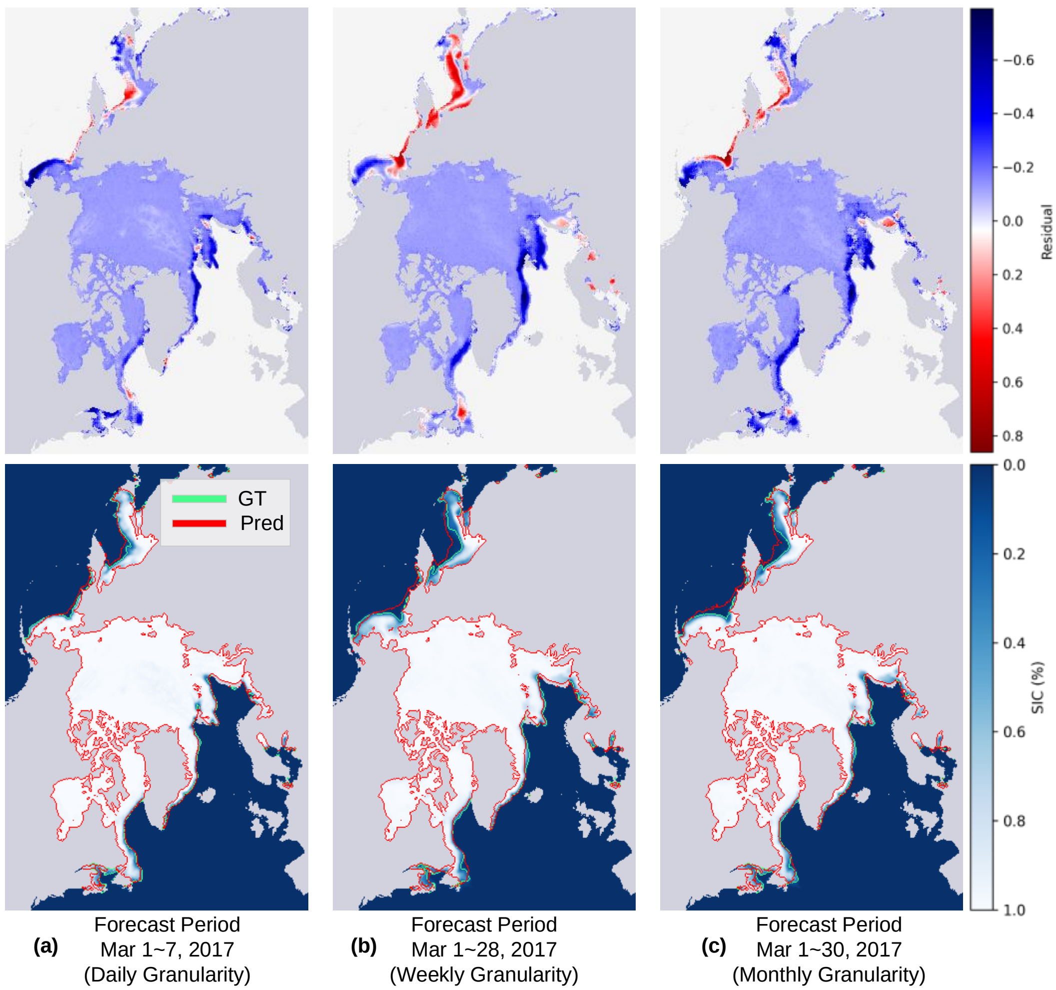


Figure 9: **Spatial residual and predicted SIE quality of Mar 2017.**

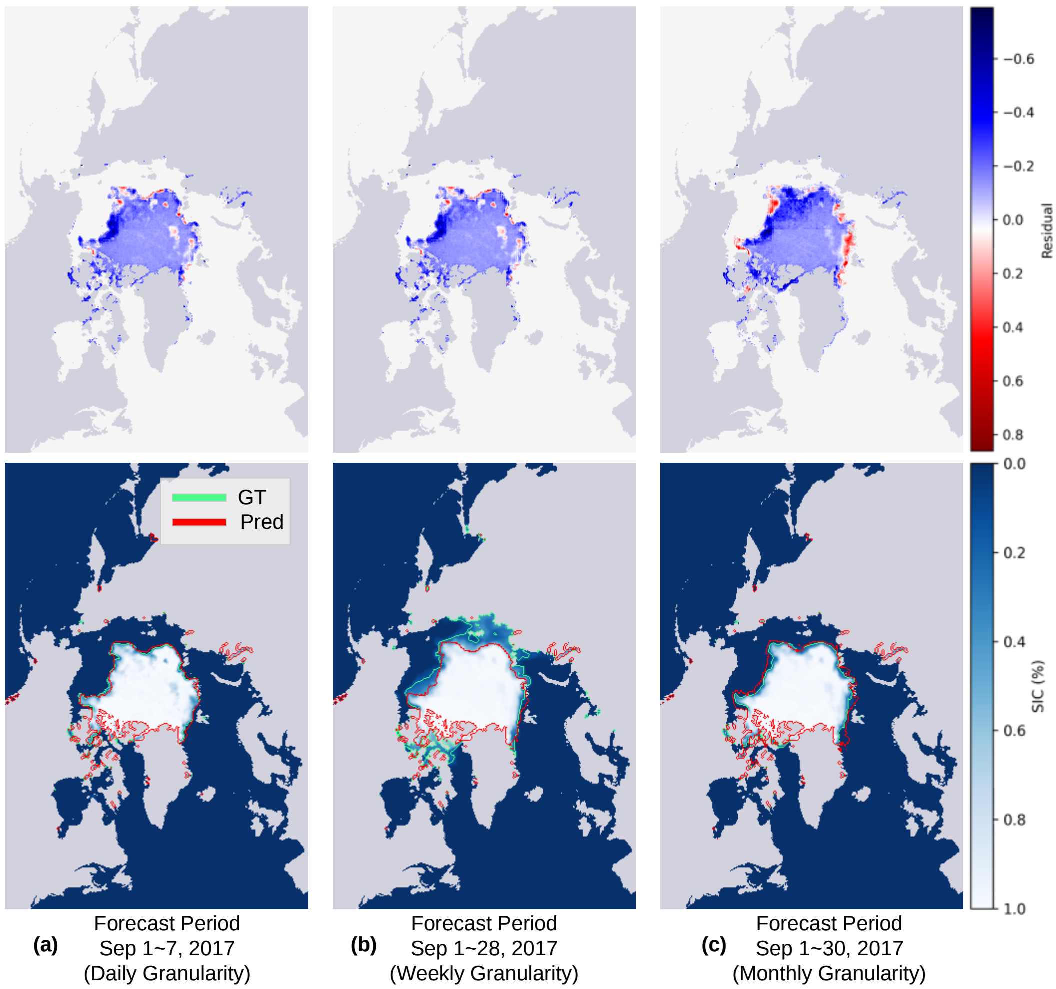


Figure 10: **Spatial residual and predicted SIE quality of Sep 2017.**

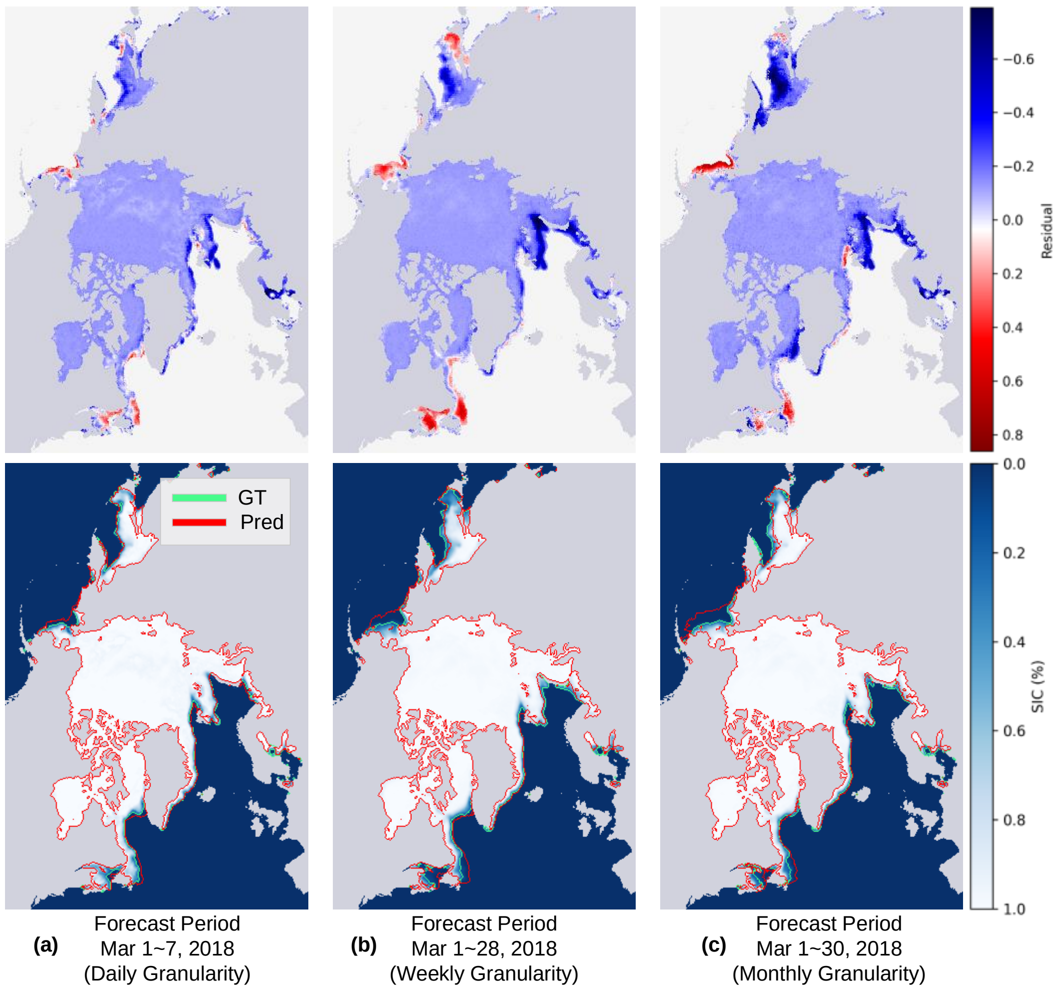


Figure 11: **Spatial residual and predicted SIE quality of Mar 2018.**

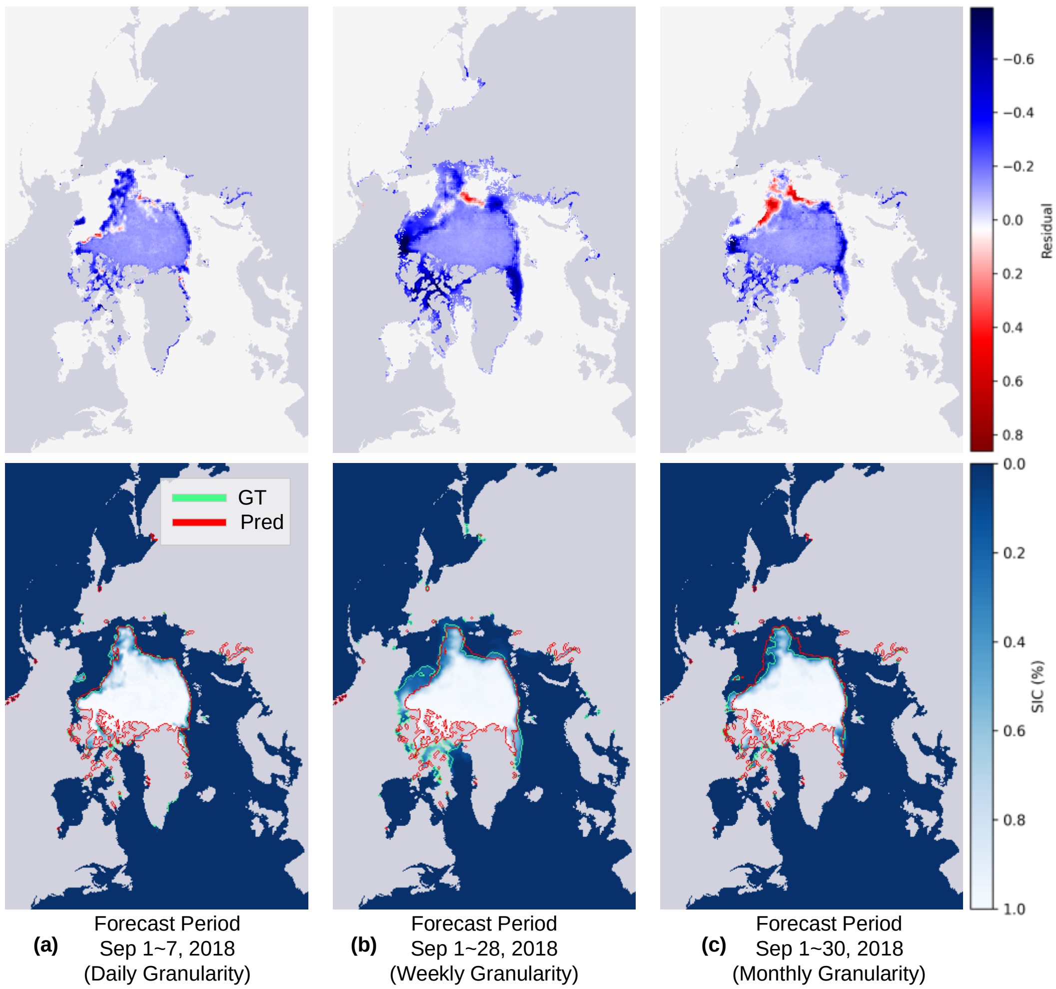


Figure 12: **Spatial residual and predicted SIE quality of Sep 2018.**

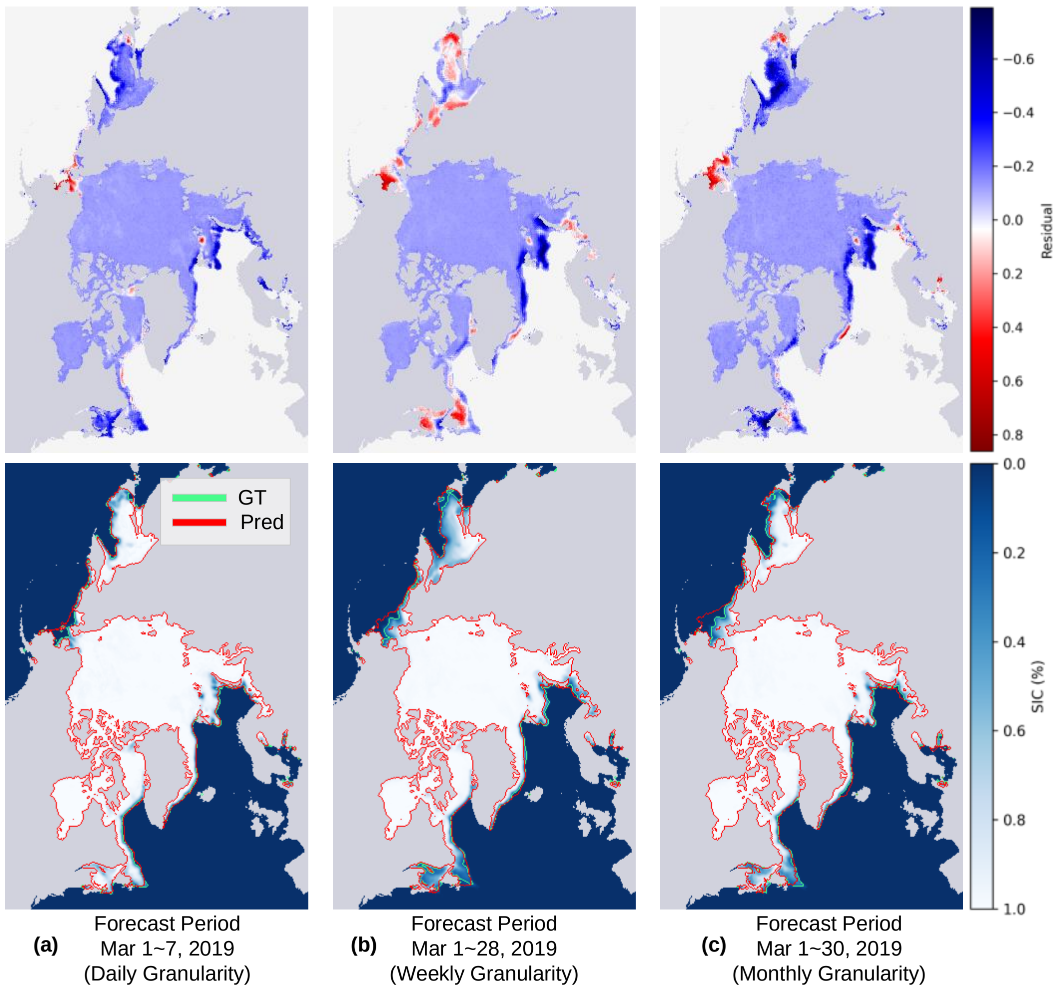


Figure 13: **Spatial residual and predicted SIE quality of Mar 2019.**

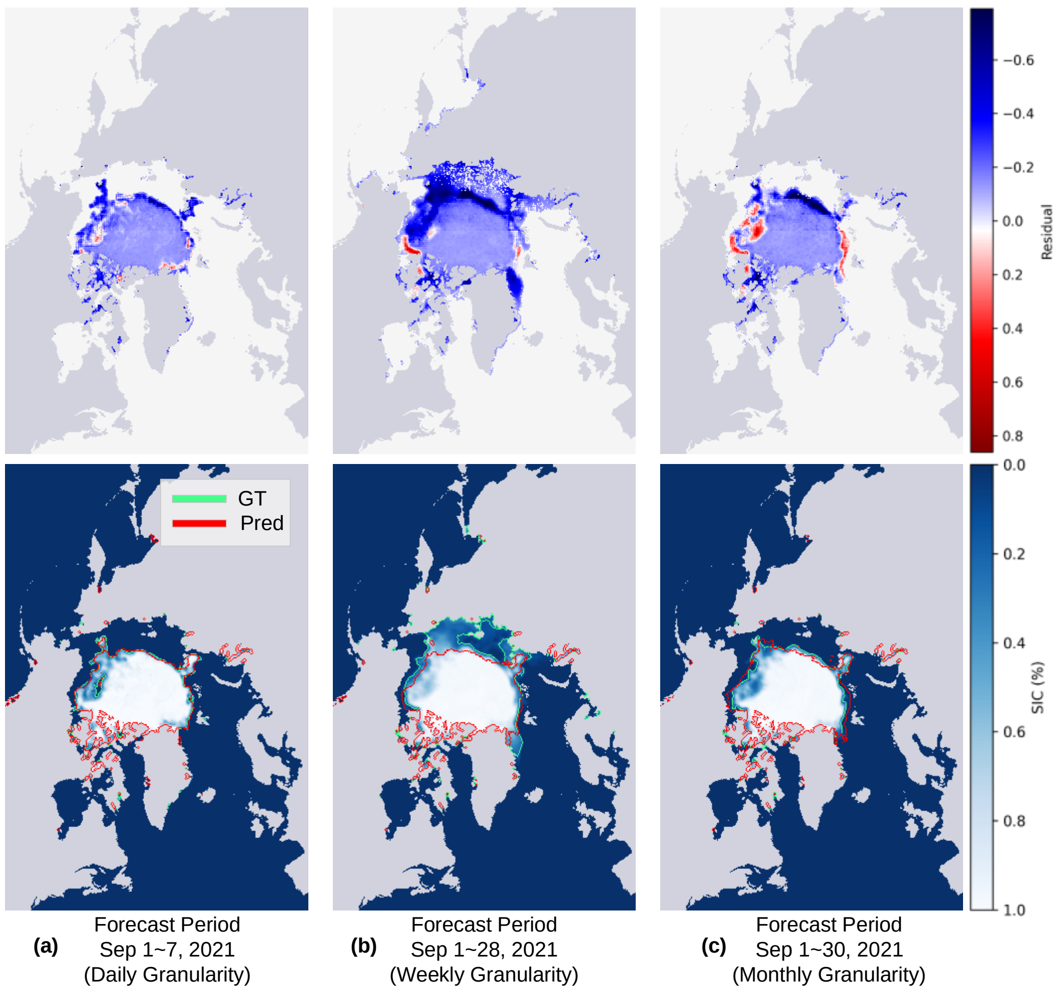


Figure 14: **Spatial residual and predicted SIE quality of Sep 2021.**

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction section of the article provide a detailed introduction to the contributions of the article.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitations of the study and suggestions for future work are elaborated in the Conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The article introduces each formula in the algorithm process and elaborates on their derivation procedures.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Section 4 and Section A.1, we present the implementation details of the experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used in the article are all publicly available. Meanwhile, the article will also make the code publicly accessible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 4 and Section A.1, we present the implementation details of the experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provides the necessary parameter settings for training and sampling, as well as training test segmentation and other related content.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix provides an introduction to the computational resources and time required for the training and sampling processes.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper fully complies with the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In the article, full citation information is provided for every point that needs referencing and included in the References section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We has organized and submitted the assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.