SEASONALITY-AWARE HYBRID CONVOLUTIONAL TRANSFORMER FOR ANTARCTIC SEA ICE FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Antarctic sea ice concentration (SIC) forecasting is an important yet challenging task due to the coexistence of complex spatial structure, long-range temporal dependencies, and strong seasonal variability. Conventional convolution-based models are effective at capturing local spatial patterns, but often have limited ability to model long-term temporal evolution. To address these challenges, we build on a hybrid convolutional transformer forecasting framework for monthly Antarctic SIC forecasting. This framework combines convolutional encoding for spatial feature extraction with space-time factorised self-attention for SIC modelling. We further introduce two season-based mechanisms: a month-aware positional encoding that injects calendar-month information into the token representation, and a seasonal temporal bias that encourages attention to periodically related historical states. Experimental results show that the proposed framework achieves better performance than convolutional and recurrent baselines as well as ECMWF's physics-based dynamical model SEAS5 under both classification and regression metrics. Ablation studies further indicate that the seasonality-aware components provide consistent additional gains in both short- and long-horizon prediction. These results demonstrate the value of combining convolutional structures, attention mechanisms, and periodic prior information for Antarctic SIC forecasting.

1 INTRODUCTION

Antarctic sea ice is a critical component of the Earth's climate system, influencing the surface energy balance and ocean-atmosphere interactions Turner et al. (2015); Parkinson (2019). Unlike the Arctic, Antarctic sea ice exhibits pronounced interannual variability and strong regional heterogeneity, with record lows and partial recoveries observed in recent decades Purich & Doddridge (2023); Josey et al. (2024). Reliable prediction of Antarctic sea ice concentration (SIC) on seasonal timescales is therefore important for climate monitoring and downstream applications such as polar logistics and marine operations. Yet Antarctic SIC forecasting remains challenging due to complex atmosphere-ocean-sea-ice interactions, sharp gradients at the ice edge, and a strong but non-stationary seasonal cycle Hobbs et al. (2016), making it a difficult spatio-temporal sequence-to-sequence learning problem.

Sea ice forecasting has traditionally relied on coupled numerical models that explicitly simulate thermodynamic and dynamic processes, including ocean-ice interactions, wind forcing, and heat exchange Mellor & Kantha (1989); Blanchard-Wrigglesworth et al. (2015); Wayand et al. (2019). Representative systems include Global Sea Ice 6.0, a CICE-based sea-ice model configuration developed for the Met Office Global Coupled model Rae et al. (2015). Seasonal Forecasting System 5 (SEAS5) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), which provides operational seasonal forecasts Johnson et al. (2019). While these approaches provide physically consistent predictions, they are computationally expensive and often require substantial effort for calibration and data assimilation, which limit their flexibility for rapid or high-frequency forecasting Liu et al. (2021).

As computationally efficient alternatives, statistical and empirical methods have also been explored for sea-ice prediction based on historical observations and climate predictors, such as multiple linear

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

regression, linear empirical models, Bayesian regression, and linear Markov models Drobot et al. (2006); Horvath et al. (2020); Yuan et al. (2016); Lindsay et al. (2008). However, many traditional statistical models rely on simplified linear or probabilistic assumptions, which may limit their ability to represent complex nonlinear interactions within the sea-ice system, particularly under extreme or rapidly changing conditions Zhu et al. (2023).

Recent advances in machine learning have introduced data-driven alternatives for sea ice prediction. Convolutional neural networks (CNNs) and their variants have demonstrated skill by learning spatial features directly from gridded SIC fields Ren et al. (2022); Andersson et al. (2021); Liu et al. (2021). Notably, IceNet Andersson et al. (2021) showed that UNet-based architectures can leverage multi-source climate inputs to achieve competitive seasonal forecasting performance. However, such 2D CNN-based approaches primarily operate as spatial pattern recognisers and lack explicit mechanisms to model long-range temporal dependencies, which are critical for multi-month prediction. To incorporate temporal dynamics, other models with spatio-temporal recurrent architectures such as Convolutional Long Short-Term Memory (ConvLSTM) and Predictive Recurrent Neural Networks (PredRNN) introduce temporal memory Shi et al. (2015); Taylor & Feng (2022); Lin et al. (2020); Wang et al. (2022). By embedding convolutional operations within recurrent units, these models preserve spatial structure while evolving temporal states. Nevertheless, their reliance on local convolutional kernels may limit their ability to capture spatial interactions across the polar domain, and training can become increasingly challenging for longer forecasting horizons Zhu et al. (2023); Liu et al. (2021).

More recently, attention-based models and Transformers have emerged as a promising alternative for modelling long-range dependencies in spatio-temporal systems Vaswani et al. (2017); Dosovitskiy et al. (2021). Self-attention mechanisms enable global information exchange across both space and time, which is particularly advantageous for capturing large-scale spatio-temporal dependencies in climate fields Gao et al. (2022); Nguyen et al. (2023). However, directly applying global self-attention to high-resolution geophysical fields remains computationally challenging due to the quadratic complexity of self-attention with respect to the number of spatial tokens. Notably, works such as TimeSformer Bertasius et al. (2021) demonstrated that factorising global self-attention into separate spatial and temporal operations drastically reduces computational complexity while maintaining state-of-the-art performance. Similar factorised and local spatio-temporal attention designs have also been explored in video vision Transformers Arnab et al. (2021); Liu et al. (2022). However, these generic video architectures are not specifically designed for geophysical forecasting and do not explicitly encode domain-specific priors such as seasonal cycles, motivating further adaptations to both manage the spatial token scale and incorporate seasonal information.

To address these challenges, we propose a convolutional transformer framework for Antarctic SIC forecasting that combines convolutional spatial feature extraction with Transformer-based spatio-temporal dependency modelling. Our main contributions are as following:

1. Convolutional transformer architecture with bottleneck attention: We develop a convolutional encoder-decoder for high-resolution spatial reconstruction, and adopt a Transformer bottleneck that performs global spatio-temporal mixing only at a compressed resolution, enabling multi-month forecasting at feasible GPU cost.
2. Month-aware Positional Encoding (MonthPE): To avoid an large 3D positional lookup table, we factorise positional and seasonal information into three distinct embeddings. This includes a standard spatial encoding, standard temporal encoding, and a Fourier-based seasonal encoding that encodes the strong seasonal periodicity of monthly sea ice.
3. Annually Temporal Attention Bias (SeasonBias): We design a divided space-time attention scheme to manage computational complexity. We augment the temporal attention logits with a learnable seasonal bias based on the circular distance of the annual cycle, directly modulating query-key interactions into seasonal states.

2 Problem Statement and Data

We frame Antarctic SIC forecasting as a spatio-temporal sequence-to-sequence prediction problem, where the model produces forecasts of future monthly SIC fields. Let the polar environment at a given time t be represented by a 2D spatial grid $X_t \in \mathbb{R}^{H \times W}$ denoting the monthly SIC input

field. H and W denote the spatial height and width. Given a historical observation window of N_{in} consecutive time steps, $\mathcal{X} = \{X_{t-(N_{in}-1)\Delta t}, X_{t-(N_{in}-2)\Delta t}, \ldots, X_t\} \in \mathbb{R}^{N_{in}\times H\times W}$, where Δt is time step size and in default, we set it as one month. The objective is to learn a mapping function $\mathcal{F}_\theta(\cdot) : \mathbb{R}^{N_{in}\times H\times W} \to \mathbb{R}^{N_{out}\times H\times W}$, parameterised by a transformer-based neural network $\mathcal{T}_\theta(\cdot) : \mathbb{R}^{N_{in}\times H\times W} \to \mathbb{R}^{H\times W}$ (see Section 3) with autoregression, to predict the future states over the next N_{out} forecast time steps:

$$\hat{\mathcal{X}} = \{\hat{X}_{t+\Delta t}, \hat{X}_{t+2\Delta t}, \ldots, \hat{X}_{t+N_{out}\Delta t}\} = \mathcal{F}_\theta(\mathcal{X}), \tag{1}$$

where θ characterises the neural network.

To train and evaluate our framework, we utilise the Sea Ice Index, Version 3 dataset provided by the National Snow and Ice Data Center (NSIDC) Fetterer et al. (2017). This dataset is derived from a continuous, multi-decadal satellite record of passive microwave observations (including SMMR, SSM/I, and SSMIS sensors), providing a highly consistent benchmark for climatological machine learning. The dataset provides monthly observations mapped to a polar stereographic projection at a spatial resolution of $25 \times 25\text{km}^2$ per pixel in 2D; see examples in Figure 1. For our network architecture, all Antarctic gridded data are uniformly resized to a fixed spatial resolution of size $H = 352$ and $W = 352$. We use data from 1978 to 2011 for training, data from 2012 to 2018 for validation, and data from 2019 to 2024 for testing.

Our models use $N_{in} = 12$ monthly averaged Antarctic SIC fields to autoregressively predict up to $N_{out} = 12$ future months. To quantitatively evaluate the performance of the proposed structure, a binary land mask is applied to restrict the quantification in forecasting domain A_Ω. Root mean squared error (RMSE) is calculated directly from the continuous SIC regression outputs. The predicted SIC fields are further divided into three classes: open water (SIC $< 15\%$), marginal ice ($15\% \leq$ SIC $\leq 80\%$), and full ice (SIC $> 80\%$) Andersson et al. (2021); Batté et al. (2020); Zampieri et al. (2018), from which three-class accuracy (Cls3 Acc), Macro-F1 score (mF1) , and F1 score are computed. We additionally report binary accuracy (Cls2 Acc) based on the 15% ice threshold:

$$\text{Binary Accuracy} = \left(1 - \frac{\text{IIEE}}{A_\Omega}\right) \times 100\%,$$

where IIEE is the integrated ice-edge error Goessling et al. (2016).

3 Methods: Convolutional Transformer Structure

To address the spatial blurring and accumulating errors in long-horizon SIC forecasting, we propose a spatio-temporal architecture that combines convolutional feature extraction with a factorised self-attention bottleneck. The convolutional encoder-decoder learns local spatial representations of SIC fields, while the temporal and spatial attention modules capture long-range dependencies and seasonally varying evolution patterns. The model produce continuous SIC forecasts for autoregressive rollout and discrete sea-ice state predictions for classification-based evaluation. We refer to Figure 1 for the overall structure. In the following subsections, we follow the top row of Figure 1 from left to right and outline the detailed design of our proposed structure.

3.1 Tokenisation of SIC Fields

Let $\mathbb{X} := \{\mathcal{X}\} \in \mathbb{R}^{N_{in}\times H\times W}$ denote a batch of historical SIC sequences. We first design a frame-wise CNN encoder for the SIC monthly field (i.e. in $H \times W$) denoted as $E(\cdot) : \mathbb{R}^{H\times W} \to \mathbb{R}^{C'\times H'\times W'}$, with a downsampling depth d to extract local spatial features and obtain a compact bottleneck representation. The encoder $E(\cdot)$ is applied to monthly SIC fields to obtain the feature space as:

$$\{E(X_t) \in \mathbb{R}^{C'\times H'\times W'}, \quad \forall X_t \in \mathbb{X}\},$$

where C' is the bottleneck channel dimension, H', W' are bottleneck spatial resolution. We flatten the spatial dimensions of $E(X_t)$ into $S = H'W'$ tokens with the set

$$U := \{\text{reshape}(E(X_t))\}_{X_t\in\mathbb{X}} \in \mathbb{R}^{N_{in}\times C'\times S}.$$

This tokenisation reduces the full spatio-temporal token length from $N_{in}HW$ at the image resolution to $N_{in}S$ at the bottleneck resolution, where $S \ll HW$. The resulting token sequence enables

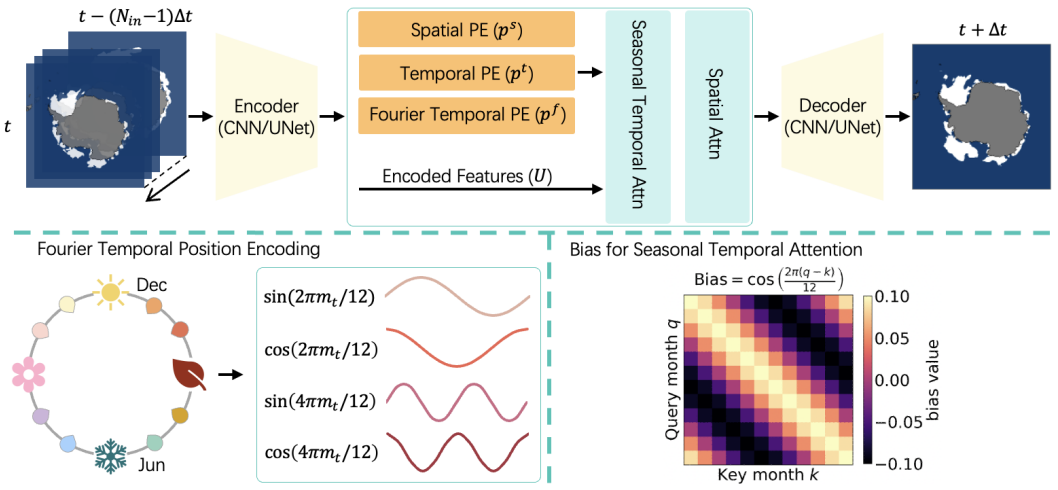

Figure 1: Overview of the proposed one-step SIC forecasting model and detailed components. Historical SIC fields are encoded into features, augmented with spatial and temporal positional encodings, processed by seasonal temporal and spatial attention, and decoded to predict SIC at $t+\Delta t$. The left lower panel shows the Fourier temporal encoding, and the right lower panel shows the seasonal bias constructed from the circular distance between query and key months.

efficient factorised spatio-temporal attention, while the convolutional encoder-decoder retains local spatial structure for fine-scale sea-ice prediction.

3.2 Month-Aware Positional Encoding

Spatio-temporal Transformer models commonly incorporate spatial and temporal positional encodings to distinguish tokens at different spatial locations and time steps Bertasius et al. (2021); Arnab et al. (2021). However, SIC forecasting exhibits strong seasonal periodicity: same calendar months across different years share similar large-scale seasonal dynamic behaviour. Thus, in addition to spatial and temporal positional encodings, we introduce month-aware positional encoding. The position encoded token can be written as

$$Z_{t,n} = U_{t,n} + \mathbf{p}^s_n + \mathbf{p}^t_t + \mathbf{p}^f_{m_t}, \qquad Z_{t,n} \in \mathbb{R}^{C'}, \tag{2}$$

where $t \in \{0, 1, \cdots, 11\}$ and $n \in \{1, 2, \cdots, S\}$ are indices of input timestep and spatial token, respectively; $m_f \in \{0, \ldots, 11\}$ denotes the calendar month associated with timestep t.

Rather than relying on a single positional representation, equation 2 separates physical priors into three distinct embeddings: (i) Let $\mathbf{P}^s \in \mathbb{R}^{S\times C'}$ represent the full spatial embedding table, and $\mathbf{p}^s_n = \mathbf{P}^s(n) \in \mathbb{R}^{C'}$ is the embedding for spatial token n. The spatial positional encoding is implemented as a learnable tensor which is added to the bottleneck features and shared across all time steps. (ii) $\mathbf{P}^t \in \mathbb{R}^{N_{in}\times C'}$ denotes the fixed sinusoidal temporal positional encoding, and $\mathbf{p}^t_t = \mathbf{P}^t(t) \in \mathbb{R}^{C'}$. (iii) $\mathbf{p}^f_{m_t} \in \mathbb{R}^{C'}$ denotes the Fourier-based calendar-month embedding, referred to MonthPE in the following. For each month index $m_t = 0, \ldots, 11$, we define a Fourier basis:

$$\Phi(m_t) = \left[\sin\left(\frac{2\pi m_t}{12}\right), \cos\left(\frac{2\pi m_t}{12}\right), \ldots, \sin\left(\frac{2\pi N_f m_t}{12}\right), \cos\left(\frac{2\pi N_f m_t}{12}\right)\right]^\top \in \mathbb{R}^{2N_f},$$

where N_f denotes the number of Fourier modes. In our implementation, we set $N_f = 4$, which allows the month encoding to retain the annual, semiannual, and several higher-order seasonal harmonics. This choice provides a balance between expressiveness and simplicity: it is sufficiently rich to represent non-sinusoidal seasonal patterns in monthly sea-ice evolution, while avoiding unnecessarily high-frequency components for a 12-month calendar. The Fourier basis is then mapped into the token space by a learnable projection W^f,

$$\mathbf{p}^f_{m_t} = W^f \Phi(m_t), \qquad W^f \in \mathbb{R}^{C'\times 2N_f}.$$

The resulting MonthPE is added to every spatial token at time step t as in equation 2, allowing the model to use an explicit 12-month periodic prior knowledge while learning how to mix different seasonal harmonics.

3.3 Factorised Spatio-Temporal Attention

Computing global self-attention over the full spatio-temporal token sequence of length $N_{in}S$ incurs an attention complexity of $\mathcal{O}((N_{in}S)^2C')$, which becomes prohibitive for high-resolution inputs (when $S \gg 1$) and long temporal contexts (when $N_{in} \gg 1$). Following Bertasius et al. (2021), we therefore adopt a divided space-time attention scheme, where attention is factorised into a temporal operator and a spatial operator applied sequentially. Specifically, temporal attention is performed independently for each spatial token, and spatial attention is then performed independently for each time step. This mechanism reduces the attention complexity to $\mathcal{O}(N_{in}^2SC' + S^2N_{in}C')$, where the first term corresponds to temporal attention and the second to spatial attention.

For notational convenience, we denote the temporal and spatial multi-head attention layers by $\mathcal{A}^t(\cdot)$ and $\mathcal{A}^s(\cdot)$, respectively, where each operator includes the corresponding attention computation and output projection. The two attention operations are applied sequentially with residual connections:

$$Z^t = Z + \mathcal{A}^t(Z), \quad Z' = Z^t + \mathcal{A}^s(Z^t),$$

where Z^t denotes the temporally representation before spatial attention layers and Z' is the output of the divided space-time attention block. Both attention operators preserve the token embedding dimension C'.

For spatial attention, each timestep is processed independently across the S spatial tokens. For each attention head, $Q^s, K^s, V^s \in \mathbb{R}^{S\times d_h}$, where d_h denotes the dimension of each attention head. Standard scaled dot-product attention is applied as

$$\text{Attn}^s(Q^s, K^s, V^s) = \text{Softmax}\left(\frac{Q^s K^{s\top}}{\sqrt{d_h}}\right) V^s.$$

The SIC latent representation Z^s from Z in equation 2 is used in the query, key and value matrices as:

$$Q^s = Z^s W_Q, \quad K^s = Z^s W_K, \quad V^s = Z^s W_V,$$

where W_Q, W_K, and W_V are learnable projection matrices.

Temporal attention is applied independently for each spatial token n. For each batch element, layer, and attention head, $Q^t, K^t, V^t \in \mathbb{R}^{N_{in}\times d_h}$. Unlike using the standard attention structure in spatial representation, to explicitly incorporate annual periodicity into temporal interactions, we introduce a seasonal bias into the temporal attention logits.

Let $q, k \in \{0, \ldots, 11\}$ denote the index associated with query timestep q and key timestep k. We define a pairwise bias matrix $b = (b_{qk})$ as

$$b_{qk} = \alpha \cos\left(\frac{2\pi(q-k)}{12}\right), \qquad \alpha \in \mathbb{R}, \tag{3}$$

where α is a learnable scalar controlling the strength of the seasonal attention bias. In particular, α is learned separately for each temporal attention layer. The magnitude of the bias b implies attention strength. A positive value indicates the attention of two nearby months, where a negative value indicates the attention of two season-opposite months. For each batch element, spatial token and attention head, a bias b is learned as in equation 3. The attention is then computed as

$$\text{Attn}^t(Q^t, K^t, V^t) = \text{Softmax}\left(\frac{Q^t K^{t\top}}{\sqrt{d_h}} + b\right) V^t,$$

Unlike positional encodings, which are injected into token representations, the proposed seasonal bias directly modulates the attention scores, thereby shaping the temporal dependency structure at the level of query-key interactions. The cosine form provides a simple and smooth periodic prior over the annual cycle, making it a natural choice for monthly geophysical data with strong seasonality. Since α is learned jointly with the rest of the network, the model can adaptively determine how strongly this seasonality-awareness should influence attention.

3.4 Decoder and Autoregressive Forecasting

To project the spatio-temporal tokens back into the physical domain, we apply a convolutional decoder, denoted as $D(\cdot)$. Following the attention bottleneck, the latent sequence is reshaped from $N_{in}SC'$ back to the spatial grid $\mathbb{R}^{N_{in}\times C'\times H'\times W'}$. The decoder maps the latent representation to the next-step SIC prediction,

$$\hat{X}_{t+\Delta t} = D(Z').$$

We consider both CNN-based and UNet-based backbone variants. For the CNN-based variant, the decoder reconstructs the output directly from the bottleneck representation without multi-scale skip connections. For the UNet-based variant, encoder features at corresponding spatial resolutions are passed to the decoder through skip connections, providing additional high-resolution spatial information during reconstruction.

Together with the encoder and spatio-temporal attention blocks described above, the complete one-step forecasting model is denoted by $\mathcal{T}_\theta$. Given the preceding N_{in} monthly SIC fields, the one-step prediction is written as

$$\hat{X}_{t+\Delta t} = \mathcal{T}_\theta\left(\{X_{t-(N_{in}-1)\Delta t}, \ldots, X_t\}\right).$$

To achieve long-horizon polar forecasting, we employ an autoregressive step. For forecast month $\ell = 1, \ldots, N_{out}$, the model is rolled out autoregressively as

$$\hat{X}_{t+\ell\Delta t} = \mathcal{T}_\theta\left(\{\tilde{X}_{t-(N_{in}-\ell)\Delta t}, \ldots, \tilde{X}_{t+(\ell-1)\Delta t}\}\right), \tag{4}$$

where

$$\tilde{X}_\tau = \begin{cases} X_\tau, & \tau \le t, \\ \hat{X}_\tau, & \tau > t. \end{cases}, \qquad \tau \in [t-(N_{in}-1)\Delta t, t+(N_{out}-1)\Delta t].$$

The overall forecasting operator $\mathcal{F}_\theta(\cdot) : \mathbb{R}^{N_{in}\times H\times W} \to \mathbb{R}^{N_{out}\times H\times W}$ is then defined as equation 4 with $\ell = 1, \ldots, N_{out}$.

Lastly, to train our neural network, we apply a loss function defined as

$$\begin{aligned} L^{\text{total}} = {} & w\big(L^{\text{mse}}(\hat{X}_{t+\Delta t}, X_{t+\Delta t}) + L^{\text{mse}}(\hat{X}_{t+2\Delta t}, X_{t+2\Delta t})\big) \\ & + (1-w)\big(L^{\text{foc}}(\hat{X}_{t+\Delta t}, X_{t+\Delta t}) + L^{\text{foc}}(\hat{X}_{t+2\Delta t}, X_{t+2\Delta t})\big), \end{aligned}$$

where L^{mse} is the standard mean squared error (MSE), L^{foc} is the Focal loss for classification with parameters to be specified in Section 4, and $w \in (0, 1)$ is the weight of the loss components. $L^{\text{mse}}(\hat{X}_{t+2\Delta t}, X_{t+2\Delta t})$ and $L^{\text{foc}}(\hat{X}_{t+2\Delta t}, X_{t+2\Delta t})$ are designed to mitigate the accumulating errors due to the autoregressive step in equation 4. Moreover, this improves the robustness and stability of the autoregression for longer time forecasting.

4 Experiments

We evaluate the proposed hybrid framework through ablation studies and comparisons with several baseline models, including CNN, UNet, ConvLSTM, PredRNN and Transformer-based variants Ronneberger et al. (2015); Shi et al. (2015); Wang et al. (2022); Bertasius et al. (2021). The experimental setup and key hyperparameters are summarised in Table 5, other settings and detailed structures are includes in Appendix A. For training, the classification branch uses Focal Loss to address class imbalance, with class weights set to $[0.2, 0.3, 0.5]$ for open water, marginal ice, and full ice, respectively. The regression branch is optimised using MSE loss. The two losses are combined with weights of 0.1 and 0.9 for the classification and regression objectives. All experiments are conducted on machines with single NVIDIA V100 GPUs.

4.1 Overall Performance Comparison

We first compare the proposed model with representative baselines to evaluate its overall forecasting performance. Figure 2 shows the performance of different deep learning models over a 12-month autoregressive forecasting horizon. As the forecast lead time increases, RMSE generally increases while classification accuracy decreases, reflecting uncertainty accumulation and error propagation

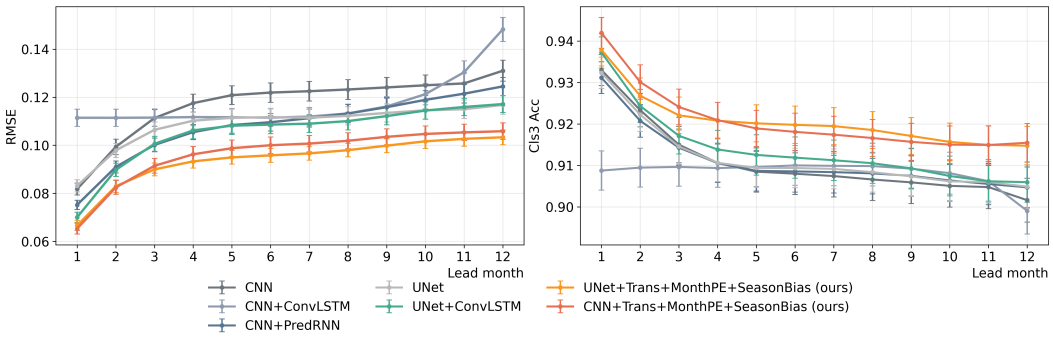

Figure 2: Comparison of multi-horizon forecasting performance across baseline and Transformer-based models over a 12-month autoregressive prediction. The left panel shows the RMSE for continuous SIC prediction, while the right panel shows the classification accuracy for categorical SIC prediction. Markers denote the mean over test samples, and error bars represent the standard error.

Table 1: Comparison with statistical, dynamical, and multi-source baselines at lead months 1, 6, and 12.

Method	RMSE@1 ↓	RMSE@6 ↓	RMSE@12 ↓	Cls3 Acc@1 ↑	Cls3 Acc@6 ↑	Cls3 Acc@12 ↑
Last-year	0.1248	0.1255	0.1292	0.9066	0.9087	0.9061
Monthly climatology	0.1056	0.1080	0.1080	0.9113	0.9111	0.9107
Anomaly persistence	0.0876	0.1261	0.1292	0.9261	0.8833	0.9061
SEAS5	0.0734	0.1279	-	0.9368	0.8972	-
ANTSIC-UNet-inspired	0.0719	0.0983	-	0.9382	**0.9215**	-
CNN+Trans+MonthPE+SeasonBias	**0.0653**	0.1000	0.1058	**0.9420**	0.9181	**0.9155**
UNet+Trans+MonthPE+SeasonBias	0.0666	**0.0958**	**0.1033**	0.9379	0.9198	0.9147

during autoregressive rollout. Nevertheless, both proposed seasonality-aware Transformer-based models with UNet and CNN backbones achieve the lowest or near-lowest RMSE and the highest or near-highest classification accuracy throughout the forecasting horizon, consistently outperforming the remaining baselines across the forecasting horizon. These results suggest that combining convolutional spatial encoding with Transformer-based temporal modelling and explicit seasonal prior information is beneficial for both continuous and categorical Antarctic SIC forecasting. A qualitative comparison of the 2023 SIC classification forecasts is provided in Appendix B.

In addition, we consider the comparison with three simple statistical methods: the same calendar month of the previous year (Last-year), monthly climatology, and anomaly persistence van den Dool (2006); Niraula & Goessling (2021). We further compare against SEAS5 over its available six-month forecast horizon and a multi-source UNet baseline inspired by IceNet and ANTSIC-UNet Andersson et al. (2021); Yang et al. (2025). The latter uses 12 months of historical SIC, recent climate variables, and 6 months of linear-trend SIC predictions. Except for Last-year, all learning-based baselines use the training split described in Section 2; all methods, including SEAS5, are evaluated over the same test period. As shown in Table 1, even without additional climate-variable inputs, the proposed models consistently outperform the statistical baselines and SEAS5 across the evaluated forecast horizons, while achieving comparable or better performance than the ANTSIC-UNet-inspired baseline. These results demonstrate the effectiveness of the proposed framework for both short- and long-range forecasting.

4.2 Ablation study at Short- and Long-Horizon Forecast Steps

To study the effect of MonthPE and SeasonBias, we compare model performance at both the first (lead time one month) and the last (lead time twelve months) autoregressive prediction steps, as reported in Table 2. This comparison allows us to assess the contribution of the seasonality-aware components in both short-range prediction and long-horizon autoregressive forecasting.

Table 2: Performance of different methods when the lead times of one and twelve months.

Backbone	Add on	Cls3 Acc	Cls3 mF1	Cls2 Acc	Cls2 F1	RMSE
Lead 1						
CNN+Trans	-	0.9371±0.0039	0.8166±0.0060	0.9727±0.0012	0.9148±0.0081	0.0679±0.0020
	+MonthPE	0.9392±0.0041	0.8216±0.0062	0.9745±0.0012	0.9209±0.0071	0.0658±0.0022
	+SeasonBias	0.9378±0.0040	0.8182±0.0062	0.9738±0.0011	0.9173±0.0078	0.0672±0.0022
	+MonthPE+SeasonBias	**0.9420**±0.0037	**0.8280**±0.0063	**0.9751**±0.0012	**0.9234**±0.0069	**0.0653**±0.0022
UNet+Trans	-	0.9353±0.0036	0.8087±0.0066	0.9703±0.0013	0.9046±0.0094	0.0714±0.0021
	+MonthPE	0.9256±0.0044	0.7827±0.0086	0.9682±0.0017	0.9011±0.0109	0.0724±0.0024
	+SeasonBias	0.9340±0.0038	0.8035±0.0067	0.9712±0.0013	0.9073±0.0092	0.0710±0.0021
	+MonthPE+SeasonBias	**0.9379**±0.0039	**0.8119**±0.0067	**0.9739**±0.0013	**0.9190**±0.0073	**0.0666**±0.0021
Lead 12						
CNN+Trans	-	0.9036±0.0047	0.6995±0.0123	0.9486±0.0020	0.8159±0.0192	0.1169±0.0035
	+MonthPE	0.9067±0.0052	0.7239±0.0089	0.9522±0.0019	0.8395±0.0151	0.1125±0.0034
	+SeasonBias	0.9098±0.0046	0.7119±0.0112	0.9511±0.0019	0.8063±0.0206	0.1194±0.0046
	+MonthPE+SeasonBias	**0.9155**±0.0046	**0.7490**±0.0086	**0.9561**±0.0019	**0.8583**±0.0129	**0.1058**±0.0034
UNet+Trans	-	0.9080±0.0049	0.7223±0.0101	0.9531±0.0020	0.8486±0.0135	0.1114±0.0038
	+MonthPE	0.9123±0.0043	**0.7510**±0.0077	0.9551±0.0015	0.8571±0.0141	**0.1026**±0.0028
	+SeasonBias	0.9089±0.0048	0.7288±0.0096	0.9538±0.0018	0.8543±0.0128	0.1063±0.0036
	+MonthPE+SeasonBias	**0.9147**±0.0047	0.7476±0.0077	**0.9576**±0.0017	**0.8615**±0.0127	0.1033±0.0031

When forecasting one month lead time, the combined use of MonthPE and SeasonBias achieves the best performance for both CNN+Trans and UNet+Trans across all reported metrics. For CNN+Trans, MonthPE alone improves all metrics over the baseline, while SeasonBias provides moderate gains; their combination further improves the results, reducing the regression RMSE from 0.0679 to 0.0653 and increasing the Cls3 Acc from 0.9371 to 0.9420. A similar trend is observed for UNet+Trans, where the combined setting achieves the best classification and regression performance, reducing RMSE from 0.0714 to 0.0666. These results indicate that the two seasonality components provide complementary short-range benefits by injecting calendar-month information and encouraging attention to seasonally related historical states.

When forecasting 12 months SIC fields, the benefit of explicit seasonality-awareness becomes more evident under long-horizon autoregressive rollout. For CNN+Trans, the combination of MonthPE and SeasonBias consistently achieves the best performance, improving Cls3 Acc from 0.9036 to 0.9155 and reducing RMSE from 0.1169 to 0.1058, indicating that the proposed components help stabilise long-range prediction. For UNet+Trans, the combined setting achieves the best Cls3 Acc, Cls2 Acc, and Cls2 F1, while MonthPE alone gives slightly better Cls3 mF1 and RMSE. Therefore, unlike the CNN-based backbone, the UNet-based backbone shows a more metric-dependent response at the final rollout step. This may be because the UNet skip connection already provides strong multi-scale spatial representations. Nevertheless, both MonthPE and SeasonBias generally improve long-horizon performance over the baseline, demonstrating the effectiveness of seasonal prior information for monthly Antarctic SIC forecasting. Additional ablation studies on SeasonBias strength and input sequence length are provided in Appendix C.

4.3 Temporal Attention

To better understand how MonthPE and SeasonBias affect the temporal behaviour of the Transformer bottleneck, we visualise the temporal attention correlations of different UNet+Trans variants in Figure 3. The baseline UNet+Trans model produces relatively diffuse attention patterns, with only weak preference for particular key months. Adding MonthPE introduces clearer month-dependent variation, suggesting that explicit calendar-month information affects how the model organises temporal memory, although the attention remains comparatively smooth. In contrast, the models with SeasonBias show much clearer diagonal and banded structures. This indicates that the temporal attention is no longer distributed uniformly over the input sequence, but is guided towards seasonally related query-key pairs. The combined MonthPE+SeasonBias model preserves this periodic structure while also allowing the attention weights to vary with the target month. This behaviour is consistent with the design of SeasonBias, which directly modulates the temporal attention logits according to the circular distance between calendar months.

These visualisations provide qualitative evidence that the proposed seasonality-aware components affect the internal temporal routing of the Transformer bottleneck. MonthPE injects seasonality

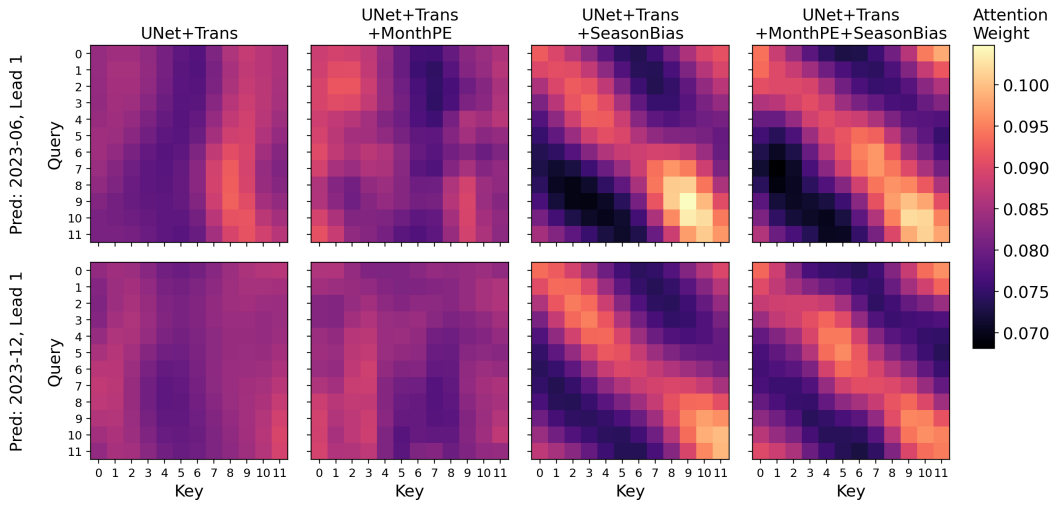

Figure 3: Temporal attention heatmaps of UNet+Trans variants for representative forecast cases. Columns correspond to the baseline model, the model with MonthPE, the model with SeasonBias, and the model with both MonthPE and SeasonBias. Rows show predictions for June and December 2023 with one month lead time.

information into the token representation, whereas SeasonBias more directly reshapes the query-key attention pattern. Together, they encourage more seasonally structured temporal dependencies, which is consistent with the improved forecasting performance observed in the quantitative results.

5 Conclusions

In this paper, we presented a hybrid convolutional transformer framework for Antarctic SIC forecasting. The proposed model combines convolutional spatial representation learning with Transformer-based spatio-temporal dependency modelling, and further incorporates seasonal information through Fourier-based positional encoding and seasonal temporal bias attention. Experimental results showed that the architecture achieves competitive forecasting performance across both classification and regression metrics. The ablation study further indicated that Fourier-based positional encoding and seasonal bias provides consistent benefits in both short- and long-range prediction. Overall, these results demonstrate the potential of hybrid convolutional and attention-based models for Antarctic SIC forecasting, especially when seasonal information is incorporated appropriately.

Our work has certain limitations. The current framework is primarily data-driven and does not explicitly use physical knowledge of the Earth system, such as thermodynamic constraints, sea-ice dynamics, or coupled ocean-atmosphere information. Prior studies suggest that incorporating physical constraints or scientific knowledge into learning-based models can improve physical consistency and robustness Karniadakis et al. (2021); Willard et al. (2022); Andersson et al. (2021). Such information may help improve the prediction of fine-scale regions, including sharp ice-edge structures and local heterogeneous patterns. In addition, since the experiments are conducted on historical Antarctic SIC data, the robustness of the model under rare extreme events and future climate regimes remains uncertain. Future work will investigate physics-guided extensions and evaluate the proposed design across broader geophysical prediction settings.

For future work, the proposed design may also be applicable to a broader range of Earth science prediction problems with strong seasonal or cyclic behaviour, and more generally to time-series forecasting tasks with known periodic patterns. Another direction is to develop more systematic interpretability analyses for attention-based SIC forecasting models. While the attention correlations in this work provide a qualitative view of the learned temporal and spatial dependencies, future studies could compare these patterns across seasons, regions, and forecast horizons to assess whether the model consistently relies on physically meaningful information. Such analyses may help identify when the model captures stable seasonal correlations.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar checking. We have not used generative AI tools for code generation, experimental design, result generation, or the development of the research ideas. All AI-assisted text was reviewed and revised by the authors to ensure the accuracy of the technical content, terminology, and citations. We take responsibility for the final content of this work.

ETHICS STATEMENT

We have reviewed the Code of Ethics and believe that the research conforms to it. The paper uses publicly available geophysical data for Antarctic sea ice concentration forecasting and does not involve human subjects, personally identifiable information, private data, or decision-making about individuals or protected groups.

REPRODUCIBILITY STATEMENT

The paper provides the information required to reproduce the main experimental results in Section 4 and Appendix A, including the data source (`https://nsidc.org/data/g02135`) and preprocessing procedures, the train/validation/test split, the 12-month input and multi-step forecasting setup, the proposed model architecture, the MonthPE and SeasonBias formulations, baseline models, loss functions, and detailed hyperparameter settings. The implementation code is included in the supplementary material, together with instructions for data preparation, training, evaluation, and reproducing the proposed method and baseline experiments. These materials provide sufficient access to the data, implementation details, and code to verify the main experimental results.

REFERENCES

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.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Lauriane Batté, Ilona Välisuo, Matthieu Chevallier, Juan C Acosta Navarro, Pablo Ortega, and Doug Smith. Summer predictions of Arctic sea ice edge in multi-model seasonal re-forecasts. *Climate Dynamics*, 54(11):5013–5029, 2020.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

E Blanchard-Wrigglesworth, RI Cullather, W Wang, J Zhang, and CM Bitz. Model forecast skill and sensitivity to initial conditions in the seasonal Sea Ice Outlook. *Geophysical Research Letters*, 42 (19):8042–8048, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Sheldon D Drobot, James A Maslanik, and Charles Fowler. A long-range forecast of Arctic summer sea-ice minimum extent. *Geophysical Research Letters*, 33(10), 2006.

F Fetterer, K Knowles, W N Meier, M Savoie, and A K Windnagel. Sea Ice Index, Version 3, 2017. URL `https://nsidc.org/data/G02135/versions/3`.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

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.

William R Hobbs, Rob Massom, Sharon Stammerjohn, Phillip Reid, Guy Williams, and Walter Meier. A review of recent changes in Southern Ocean sea ice, their drivers and forcings. *Global and Planetary Change*, 143:228–250, 2016.

Sean Horvath, Julienne Stroeve, Balaji Rajagopalan, and William Kleiber. A Bayesian logistic regression for probabilistic forecasts of the minimum September Arctic sea ice cover. *Earth and Space Science*, 7(10):e2020EA001176, 2020.

Stephanie J. Johnson, Timothy N. Stockdale, Laura Ferranti, Magdalena A. Balmaseda, Franco Molteni, Linus Magnusson, Steffen Tietsche, Damien Decremer, Antje Weisheimer, Gianpaolo Balsamo, Sarah P. E. Keeley, Kristian Mogensen, Hao Zuo, and Beatriz M. Monge-Sanz. SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117, 2019. doi: 10.5194/gmd-12-1087-2019.

Simon A Josey, Andrew JS Meijers, Adam T Blaker, Jeremy P Grist, Jenny Mecking, and Holly C Ayres. Record-low Antarctic sea ice in 2023 increased ocean heat loss and storms. *Nature*, 636 (8043):635–639, 2024.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. Self-attention ConvLSTM for Spatiotemporal Prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11531–11538, 2020.

RW Lindsay, Jinlun Zhang, AJ Schweiger, and MA Steele. Seasonal predictions of ice extent in the Arctic Ocean. *Journal of Geophysical Research: Oceans*, 113(C2), 2008.

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.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.

George L Mellor and Lakshmi Kantha. An Ice-Ocean Coupled Model. *Journal of Geophysical Research: Oceans*, 94(C8):10937–10954, 1989.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A Foundation Model for Weather and Climate. In *1st Workshop on the Synergy of Scientific and Machine Learning Modeling @ ICML2023*, 2023. URL https://openreview.net/forum?id=5CZyPMgjUu.

Bimochan Niraula and Helge F. Goessling. Spatial Damped Anomaly Persistence of the Sea Ice Edge as a Benchmark for Dynamical Forecast Systems. *Journal of Geophysical Research: Oceans*, 126(12):e2021JC017784, 2021. doi: https://doi.org/10.1029/2021JC017784. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JC017784. e2021JC017784 2021JC017784.

Claire L Parkinson. A 40-y record reveals gradual Antarctic sea ice increases followed by decreases at rates far exceeding the rates seen in the Arctic. *Proceedings of the National Academy of Sciences*, 116(29):14414–14423, 2019.

Ariaan Purich and Edward W Doddridge. Record low Antarctic sea ice coverage indicates a new sea ice state. *Communications Earth & Environment*, 4(1):314, 2023.

JGL Rae, HT Hewitt, AB Keen, JK Ridley, AE West, CM Harris, Elizabeth Clare Hunke, and DN Walters. Development of the global sea ice 6.0 CICE configuration for the Met Office global coupled model. *Geoscientific Model Development*, 8(7):2221–2230, 2015.

Yibin Ren, Xiaofeng Li, and Wenhao Zhang. A data-driven deep learning model for weekly sea ice concentration prediction of the pan-arctic during the melting season. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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.

John Taylor and Ming Feng. A deep learning model for forecasting global monthly mean sea surface temperature anomalies. *Frontiers in Climate*, 4:932932, 2022.

John Turner, J Scott Hosking, Thomas J Bracegirdle, Gareth J Marshall, and Tony Phillips. Recent changes in Antarctic sea ice. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2045), 2015.

Huug van den Dool. Methods in Short-Term Climate Prediction. In *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 12 2006. ISBN 9780199202782. doi: 10.1093/oso/9780199202782.003.0015. URL https://doi.org/10.1093/oso/9780199202782.003.0015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022.

NE Wayand, CM Bitz, and E Blanchard-Wrigglesworth. A Year-Round Subseasonal-to-Seasonal Sea Ice Prediction Portal. *Geophysical Research Letters*, 46(6):3298–3307, 2019.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37, 2022.

Ziying Yang, Jiping Liu, Mirong Song, Yongyun Hu, Qinghua Yang, Ke Fan, Rune Grand Graversen, and Lu Zhou. Extended seasonal prediction of Antarctic sea ice concentration using ANTSIC-UNet. *The Cryosphere*, 19(12):6381–6402, 2025. doi: 10.5194/tc-19-6381-2025. URL https://tc.copernicus.org/articles/19/6381/2025/.

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.

Lorenzo Zampieri, Helge F Goessling, and Thomas Jung. Bright prospects for Arctic sea ice prediction on subseasonal time scales. *Geophysical Research Letters*, 45(18):9731–9738, 2018.

Yilin Zhu, Mengjiao Qin, Panxi Dai, Sensen Wu, Zhiyi Fu, Zhende Chen, Laifu Zhang, Yuanyuan Wang, and Zhenhong Du. Deep learning-based seasonal forecast of sea ice considering atmospheric conditions. *Journal of Geophysical Research: Atmospheres*, 128(24):e2023JD039521, 2023.

A NETWORK STRUCTURE AND PARAMETERS

Additional implementation details of the network architectures are provided as follows. The main text focuses on the proposed seasonality-aware components, while this appendix summarises the corresponding backbone structures, bottleneck designs, and parameter counts for reproducibility. All models use the same input-output setting and are evaluated under the same training and testing protocol unless otherwise stated.

Table 3 and 4 summarise the detailed structure of the proposed convolutional transformer model. The convolutional encoder first extracts local spatial features and reduces the spatial resolution before the Transformer bottleneck is applied. At the bottleneck, the features are converted into spatio-temporal tokens, where positional encodings and factorised self-attention are used to model temporal and spatial dependencies. The decoder then upsamples the features back to the original spatial resolution and produces both regression and classification outputs.

Table 5 summarises the main forecasting, model, seasonal-prior, training, and loss settings. All compared models use the same input resolution, historical input window, forecast horizon, loss functions, and training protocol unless otherwise specified. Table 6 further compares the bottleneck modules used in different baselines. The pure CNN and UNet baselines rely on convolutional bottlenecks without a temporal modelling module. The ConvLSTM and PredRNN variants introduce recurrent temporal modelling at the bottleneck feature level. The Transformer-based variants reshape the bottleneck features into spatio-temporal tokens and apply factorised temporal and spatial self-attention. The main architectural difference among all models lies in the bottleneck module, while other settings and training strategy are kept consistent. Since the attention modules are applied only after spatial downsampling at the bottleneck resolution, all experiments can be run on a single NVIDIA V100 GPU with gradient checkpointing, without requiring model parallelism.

Table 3: Detailed architecture of the proposed UNet-Transformer model with MonthPE and SeasonBias. Here $N_{in} = 12$, $H = W = 352$, $S = 44 \times 44$, and $C_{in} = 256$. The batch dimension is omitted for readability.

Block	Layer	Resolution	Channels
Input	–	352×352	1
Encoder Stem	Conv 3×3	352×352	$1 \to 32$
	ReLU	352×352	32
	Conv 3×3	352×352	32
	ReLU	352×352	32
Encoder Down 1	MaxPool 2×2	$352 \times 352 \to 176 \times 176$	32
	Conv 3×3	176×176	$32 \to 64$
	ReLU	176×176	64
	Conv 3×3	176×176	64
	ReLU	176×176	64
Encoder Down 2	MaxPool 2×2	$176 \times 176 \to 88 \times 88$	64
	Conv 3×3	88×88	$64 \to 128$
	ReLU	88×88	128
	Conv 3×3	88×88	128
	ReLU	88×88	128
Encoder Down 3	MaxPool 2×2	$88 \times 88 \to 44 \times 44$	128
	Conv 3×3	44×44	$128 \to 256$
	ReLU	44×44	256
	Conv 3×3	44×44	256
	ReLU	44×44	256
Bottleneck Embedding	Learnable spatial PE	44×44	256
	Flatten to tokens	$44 \times 44 \to S = 1936$	256
	Relative time PE	$N_{in} \times S$	256
	MonthPE	$N_{in} \times S$	256
Transformer Block $\times 2$	Temporal reshape	$N_{in} \times S \to S \times N_{in}$	256
	LayerNorm	$S \times N_{in}$	256
	Temporal MHA	$S \times N_{in}$	256
	Residual	$S \times N_{in}$	256
	Spatial reshape	$S \times N_{in} \to N_{in} \times S$	256
	LayerNorm	$N_{in} \times S$	256
	Spatial MHA	$N_{in} \times S$	256
	Residual	$N_{in} \times S$	256
	FFN	$N_{in} \times S$	$256 \to 1024 \to 256$
	Residual	$N_{in} \times S$	256
Decoder Up 1	ConvTranspose 2×2	$44 \times 44 \to 88 \times 88$	$256 \to 128$
	Concat skip	88×88	$128 + 128 = 256$
	Conv 3×3	88×88	$256 \to 128$
	ReLU	88×88	128
	Conv 3×3	88×88	128
	ReLU	88×88	128
Decoder Up 2	ConvTranspose 2×2	$88 \times 88 \to 176 \times 176$	$128 \to 64$
	Concat skip	176×176	$64 + 64 = 128$
	Conv 3×3	176×176	$128 \to 64$
	ReLU	176×176	64
	Conv 3×3	176×176	64
	ReLU	176×176	64
Decoder Up 3	ConvTranspose 2×2	$176 \times 176 \to 352 \times 352$	$64 \to 32$
	Concat skip	352×352	$32 + 32 = 64$
	Conv 3×3	352×352	$64 \to 32$
	ReLU	352×352	32
	Conv 3×3	352×352	32
	ReLU	352×352	32
Output Heads	Classification head (1×1 Conv)	352×352	$32 \to C_{\text{cls}}$
	Regression head (1×1 Conv)	352×352	$32 \to 1$
	Classification temporal weighting	352×352	C_{cls}
	Regression temporal weighting + Sigmoid	352×352	1

Table 4: Detailed architecture of the proposed CNN-Transformer model with MonthPE and SeasonBias. Here $N_{in} = 12$, $H = W = 352$, $S = 44 \times 44$, and $C_{in} = 256$. The batch dimension is omitted for readability.

Block	Layer	Resolution	Channels
Input	–	352×352	1
Encoder Stem	Conv 3×3	352×352	$1 \to 32$
	ReLU	352×352	32
	Conv 3×3	352×352	32
	ReLU	352×352	32
Encoder Down 1	MaxPool 2×2	$352 \times 352 \to 176 \times 176$	32
	Conv 3×3	176×176	$32 \to 64$
	ReLU	176×176	64
	Conv 3×3	176×176	64
	ReLU	176×176	64
Encoder Down 2	MaxPool 2×2	$176 \times 176 \to 88 \times 88$	64
	Conv 3×3	88×88	$64 \to 128$
	ReLU	88×88	128
	Conv 3×3	88×88	128
	ReLU	88×88	128
Encoder Down 3	MaxPool 2×2	$88 \times 88 \to 44 \times 44$	128
	Conv 3×3	44×44	$128 \to 256$
	ReLU	44×44	256
	Conv 3×3	44×44	256
	ReLU	44×44	256
Bottleneck Embedding	Learnable spatial PE	44×44	256
	Flatten to tokens	$44 \times 44 \to S = 1936$	256
	Relative time PE	$N_{in} \times S$	256
	MonthPE	$N_{in} \times S$	256
Transformer Block $\times 2$	Temporal reshape	$N_{in} \times S \to S \times N_{in}$	256
	LayerNorm	$S \times N_{in}$	256
	Temporal MHA + SeasonBias	$S \times N_{in}$	256
	Residual	$S \times N_{in}$	256
	Spatial reshape	$S \times N_{in} \to N_{in} \times S$	256
	LayerNorm	$N_{in} \times S$	256
	Spatial MHA	$N_{in} \times S$	256
	Residual	$N_{in} \times S$	256
	FFN	$N_{in} \times S$	$256 \to 1024 \to 256$
	Residual	$N_{in} \times S$	256
CNN Decoder Up 1	ConvTranspose 2×2	$44 \times 44 \to 88 \times 88$	$256 \to 128$
	Conv 3×3	88×88	$128 \to 128$
	ReLU	88×88	128
	Conv 3×3	88×88	128
	ReLU	88×88	128
CNN Decoder Up 2	ConvTranspose 2×2	$88 \times 88 \to 176 \times 176$	$128 \to 64$
	Conv 3×3	176×176	$64 \to 64$
	ReLU	176×176	64
	Conv 3×3	176×176	64
	ReLU	176×176	64
CNN Decoder Up 3	ConvTranspose 2×2	$176 \times 176 \to 352 \times 352$	$64 \to 32$
	Conv 3×3	352×352	$32 \to 32$
	ReLU	352×352	32
	Conv 3×3	352×352	32
	ReLU	352×352	32
Output Heads	Classification head (1×1 Conv)	352×352	$32 \to C_{\text{cls}}$
	Regression head (1×1 Conv)	352×352	$32 \to 1$
	Classification temporal weighting	352×352	C_{cls}
	Regression temporal weighting + Sigmoid	352×352	1

Table 5: Hyperparameters and experimental setup.

Category	Hyperparameter	Value
Forecasting	Input resolution	352×352
	Historical input window (N_{in})	12 months
	Forecast horizon (K)	12 months
	Output classes	3(Cls) and 1(Reg)
Model	Spatial depth (d)	3
	Temperal Transformer layers	2
	Spacial Transformer layers	2
	Attention heads (h)	8
	MLP ratio	4
Seasonality-aware components	MonthPE Fourier modes (N_f)	4
	SeasonBias initialization (α_0)	0.1
	Attention dropout	0.0
Training	Optimiser	Adam
	Learning rate	1×10^{-5}
	Batch size	1
	Training epochs	200
Loss	Classification loss	Focal loss
	Regression loss	MSE loss
	Loss weights [Cls, Reg]	[0.1, 0.9]

Table 6: Implementation details of the bottleneck modules used in the compared models. Here $N_{in} = 12$, $S = 44 \times 44 = 1936$, and $C = 256$. '-' indicates that no explicit temporal bottleneck module is used.

Model	Bottleneck input	Bottleneck module	Implementation details
CNN	$N_{in} \times C \times 44 \times 44$	-	standard convolutional bottleneck
UNet	$N_{in} \times C \times 44 \times 44$	-	standard UNet bottleneck with skip connections
CNN+ConvLSTM	$N_{in} \times C \times 44 \times 44$	ConvLSTM	ConvLSTM down/up bottleneck hidden channels $= 256$ kernel size $= 3 \times 3$
UNet+ConvLSTM	$N_{in} \times C \times 44 \times 44$	ConvLSTM	ConvLSTM down/up bottleneck hidden channels $= 256$ kernel size $= 3 \times 3$
CNN+PredRNN	$N_{in} \times C \times 44 \times 44$	PredRNN	PredRNN bottleneck hidden channels $= 256$
CNN+Trans+MonthPE+SeasonBias	$N_{in} \times S \times C$	Factorised Transformer	Position Encoder temporal MHA, heads $= 8$ spatial MHA, heads $= 8$
UNet+Trans+MonthPE+SeasonBias	$N_{in} \times S \times C$	Factorised Transformer	Position Encoder temporal MHA, heads $= 8$ spatial MHA, heads $= 8$

B Qualitative Comparison

To further examine the spatial characteristics of the forecasts, Figure 4 provides a qualitative comparison of SIC classification for the 2023 prediction period. All models capture the dominant seasonal expansion and retreat of Antarctic sea ice, but visible differences appear near the marginal ice zone, where classification is more difficult due to stronger spatial variability. The recurrent and convolutional baselines tend to produce smoother or less accurate ice-edge structures, whereas the Transformer-based models with MonthPE and SeasonBias better preserve the seasonal evolution of the ice boundary. This visual comparison is consistent with the quantitative results in Section 4,Figure 2, showing that explicit seasonality-awareness improve the model's ability to represent both the temporal and spatial structure of SIC evolution.

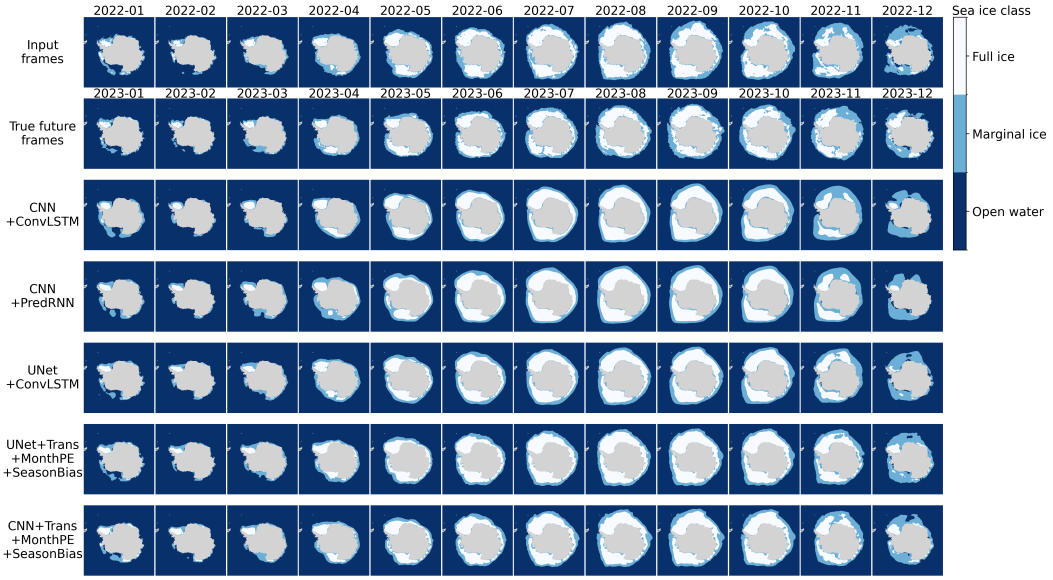

Figure 4: Qualitative comparison of SIC classification forecasts for 2023. The first row shows the 12-month input sequence from 2022, and the second row shows the corresponding ground-truth SIC classes for 2023. The remaining rows compare the predicted SIC classes generated by different forecasting models. Grey regions indicate masked land.

C Additional ablation studies

We perform additional ablation studies on SeasonBias strength. Table 7 reports the twelve lead months forecasting performance of the UNet+Trans backbone with MonthPE under different initial SeasonBias strengths. The results show that the model is relatively robust to the choice of α, with only moderate variations across the tested range. Among the tested settings, $\alpha = 0.12$ achieves the best overall performance, obtaining the highest Cls3 Acc, Cls3 mF1, Cls2 Acc, Cls2 F1, and the lowest regression RMSE. This suggests that a moderate seasonal bias provides the most effective prior for long-range forecasting, while a too weak or too strong bias may slightly reduce either classification or regression performance.

Table 8 evaluates the effect of increasing the input window from 12 to 24 months. Although the model maintains reasonable forecasting performance at both one lead month and twelve lead months, the longer input window does not provide a clear overall improvement over the 12-month input setting. This suggests that simply extending the historical context is not necessarily beneficial for SIC forecasting, and that the model may require more effective mechanisms to select or weight useful temporal information from longer sequences.

Table 7: Twelve lead months forecasting performance under different SeasonBias strengths (mean ± SE over 48 test samples) on UNet+Trans backbone with MonthPE.

SeasonBias	Cls3 Acc	Cls3 mF1	Cls2 Acc	Cls2 F1	RMSE
$\alpha = 0.04$	0.9068±0.0050	0.7178±0.0090	0.9518±0.0018	0.8420±0.0136	0.1160±0.0038
$\alpha = 0.08$	0.9066±0.0052	0.7091±0.0104	0.9536±0.0020	0.8494±0.0136	0.1093±0.0033
$\alpha = 0.12$	**0.9117**±0.0047	**0.7402**±0.0082	**0.9538**±0.0020	**0.8538**±0.0127	**0.1075**±0.0032
$\alpha = 0.16$	0.9093±0.0048	0.7133±0.0102	0.9527±0.0019	0.8365±0.0154	0.1184±0.0039
$\alpha = 0.20$	0.9098±0.0048	0.7082±0.0122	0.9542±0.0020	0.8514±0.0135	0.1108±0.0032

Table 8: Forecast performance at one lead month and twelve lead months using 24 input months.

Lead	Cls3 Acc	Cls3 mF1	Cls2 Acc	Cls2 F1	RMSE
1	0.9355 ± 0.0094	0.8080 ± 0.0128	0.9726 ± 0.0032	0.9131 ± 0.0167	0.0661 ± 0.0046
12	0.9146 ± 0.0105	0.7527 ± 0.0188	0.9569 ± 0.0043	0.8691 ± 0.0244	0.0988 ± 0.0075