**General comments:**

In summary, this research aims to develop a transformer-based U-net model to improve prediction skill of Arctic sea ice concentration at seasonal lead times. The primary novel contribution of this work is the use of a state-of-the-art deep learning architecture with a custom domain-specific loss function as well as including spring sea ice thickness as an input predictor to overcome the common spring predictability barrier (SPB) faced by previous numerical and DL-based models that decrease prediction skill of predictions made before May.

The authors choose to use a transformer based architecture to capture spatiotemporal patterns in the input data. The model inputs include multiple sea-ice related variables over 3-6 months and outputs the sea ice concentration of the next six months, specifically looking at predictions for June-Sept at 6 month lead times. The authors develop a novel loss function that combines both a standard DL loss function (mean squared error) and a domain-specific NIEE loss which accounts for spatial similarity in model predictions and ground truth. The model follows a standard DL training procedure and includes three years of unseen data to test their model performance. The authors evaluate their model using standard DL performance metrics along with Binary Accuracy (BACC) which accounts for accurate spatial distributions.

The authors succeed in improving prediction skill when compared to the benchmark Persistence model and dynamical ECMWF model. For Sept ice prediction specifically, the ECMWF model performs better at smaller lead times but the author's model shows marked performance improvement at a lead time of 3-6 months, thus overcoming the spring predictability barrier. The authors perform necessary ablation tests to investigate the role of sea ice thickness in their model. They show a model that includes sea ice thickness as input outperforms a model without sea ice thickness at seasonal lead times of 3-6 months. When testing their model on unseen data, the authors use an ensemble model approach by averaging the results of 20 separately trained models. The authors also test their model against a previous state of the art deep learning CNN based approach, IceNet. They show their model produces better ACC scores at longer lead times. Their model produced lower BACC scores but can capture more individualized/local or extreme characteristics in comparison to IceNet which tends to produce smoothed out results.

Overall, this manuscript provides a significant improvement to the modeling of sea ice concentration by providing a novel approach to increase prediction skill at seasonal lead times to overcome the SPB. The authors clearly state their motivations for the project and outline their novel contributions. The authors discuss previous modeling approaches in the field and clearly showcase their significant scientific results where their model outperforms previous approaches. The contributions of this paper is twofold, one in utilizing a novel transformer based approach that the authors state has not been previously used in this field and second in highlighting the impact of spring sea ice thickness on improving prediction accuracies, revealing potential scientific insights that need to be studied further. Below are a few suggestions to improve reproducibility and presentation quality.
**Response:** Thanks for the comment.

**Specific comments:**

**Comment 1**: In line 52, the authors state that experiments show the reason for decrease in prediction skill before spring is due to ice motion and growth in the winter. It is unclear which experiments in the literature the authors are referring to. Including a citation here would help justify this statement.

**Response:** Thanks for the comment. We cite the following reference in this sentence:

Bushuk, M., Winton, M., Bonan, D. B., Blanchard-Wrigglesworth, E., and Delworth, T. L.: A mechanism for the Arctic sea ice spring predictability barrier, Geophys Res Lett, 47, https://doi.org/10.1029/2020GL088335, 2020.

**Comment 2**: In line 79, a citation for the IceNet model (Andersson et al 2021) is missing. Since working with the IceNet model is a significant portion of the manuscript, perhaps including more information about the IceNet model and how it differentiates from the author's transformer based model would provide more context to the reader in this section.

**Response:** Thanks for the comment. We added the citation for IceNet in line 79. A brief introduction about IceNet has also been added. The revised sentences are as follows:

Finally, we compare our $\text{SICNet}_{\text{season}}$ model with the published deep learning model IceNet (Andersson et al., 2021). IceNet is a probability prediction model for Arctic SIE based on convolutional neural network (CNN) units and the U-Net architecture. It achieved state-of-the-art performance in predicting the SIE for six months (Andersson et al., 2021). Therefore, we chose IceNet as a comparison model.

Besides, more information about the IceNet model, such as the model's inputs and outputs, was presented in Section 4.7, lines 263-267:

The IceNet is a seasonal sea ice prediction model that performs state-of-the-art SIE prediction (Andersson et al., 2021). It is a CNN-based classification model, and it outputs the probability of three classes: open water (SIC≤15%), marginal ice (15% < SIC < 80%), and full ice (SIC≥80%). Differently, our $\text{SICNet}_{\text{season}}$ outputs the 0-100% range SIC values. The IceNet's inputs consist of 50 monthly mean variables, including SIC, 11 climate variables, statistical SIC forecasts, and metadata.

**Comment 3**: In lines 108-109, the authors state that multiple experiments were conducted to determine the length of the model input features. It is unclear what experiments or methods the authors tried. Was this determined by manual tuning using their cross-validation strategy or did the authors employ an automated grid-search type strategy. The authors can also include if domain knowledge or previous literature informed the choice of input lengths of the data.

**Response:** Thanks for the comment. We determine the length of input factors by combining domain knowledge and manual tuning experiments. The main domain knowledge we considered is the sea ice reemergence mechanism. The spring-fall reemergence occurs between pairs of months where the ice edge is in the same position, such as in May and December (Blanchard-Wrigglesworth et al., 2011; Day et al., 2014). The spring sea ice anomaly is positively correlated with fall sea ice anomalies, and there is also a weaker reemergence between fall sea ice anomalies and anomalies the following spring (Bushuk et al., 2015). Therefore, we set the initial input length of the SIC/SIT/SIC anomaly as six months. Then, we change the length manually to fine-tune the deep learning model to find the best-matched length for each factor. The SIC climatology of the target months provides an essential mean state of the prediction SIC, so we input it into the model. It can also represent the month number signal that IceNet has considered. We explain more details about this issue in the revised manuscript, and the new revision is as follows:

The input for $\text{SICNet}_{\text{season}}$ is a 96×96×18 SIC and SIT sequence, composed of SIT of the last three months, SIC of the last six months, SIC anomaly of the last three months, and SIC climatology of the six target months (Fig. 1a). We determine the length of input factors by combining domain knowledge and manual tuning experiments. The primary domain knowledge we considered is the spring-fall reemergence mechanism. It occurs between pairs of months where the ice edge is in the same position, such as in May and December (Blanchard-Wrigglesworth et al., 2011; Day et al., 2014). The spring sea ice anomaly is positively correlated with fall sea ice anomalies, and there is also a weaker reemergence between fall sea ice anomalies and anomalies the following spring (Bushuk et al., 2015). Therefore, we set the initial input length of the SIC/SIT/SIC anomaly as six months. We change the input length manually (from six to one in step one) to fine-tune the deep learning model to find the best-matched length for each factor. The SIC climatology of the target months provides an essential mean state of the prediction SIC. It represents the monthly cycle signal that IceNet has considered.

[1] Blanchard-Wrigglesworth, E., Armour, K. C., Bitz, C. M., and Deweaver, E.: Persistence and inherent predictability of arctic sea ice in a GCM ensemble and observations, J Clim, 24, 231–250, https://doi.org/10.1175/2010JCLI3775.1, 2011.

[2] Day, J. J., Tietsche, S., and Hawkins, E.: Pan-arctic and regional sea ice predictability: Initialization month dependence, J Clim, 27, 4371–4390, https://doi.org/10.1175/JCLI-D-13-00614.1, 2014.

[3] Bushuk, M., Giannakis, D., and Majda, A. J.: Arctic Sea Ice Reemergence: The Role of Large-Scale Oceanic and Atmospheric Variability*, https://doi.org/10.1175/JCLI-D-14-00354.s1, 2015.

**Comment 4**: In section 4.7, when comparing the transformer based model to the CNN-based IceNet model, the authors state they used identical training and testing settings to perform fair comparisons. It is unclear whether the authors used the same 20 trained ensemble model approach they had used for their transformer model for the IceNet model. If so, specifying whether they used an ensemble approach or singular-model approach for IceNet would clarify this for the reader. If the authors did not use a similar ensemble approach, the authors should justify this choice.

**Response:** Thanks for the comment. We are sorry for the confusion. We did not use an ensemble approach for both $\text{SICNet}_{\text{season}}$ and IceNet. The training procedure is a leave-one-year-out strategy for the 20 testing years (2000-2019). For example, if the testing year is 2019, the training set is data from 1979-2018, and the testing data is 2019. Then, the testing year moves to 2018, and the corresponding training set is data from 1979-2017 and 2019. The model is trained 3 times for each training-testing pair to eliminate randomness, and the prediction for each testing year is the mean value of the three trained models. We explain the leave-one-year-out training procedure in Section 4.1. Further, we clarify the training strategy in the revision:

The training and testing settings of IceNet are the same as those of $\text{SICNet}_{\text{season}}$. The IceNet is trained using the same leave-one-year-out strategy as the $\text{SICNet}_{\text{season}}$. For example, if the testing year is

2019, the training set is data from 1979-2018, and the testing data is 2019. Then, the testing data moves to 2018; the remaining data (1979-2017, 2019) is the training set. For each training/testing pair, the model is trained three times to eliminate randomness, and the final prediction for testing data is the mean value of the three models.

**Comment 5**: In line 269, it is unclear how the authors transformed the IceNet output to match the continuous scale (for e.g restructuring only the final output layer). Having this additional context would help with reproducibility of this experiment.

**Response:** Thanks for the comment. We reconstruct IceNet's output layer by replacing the softmax with the sigmoid activation function. The sigmoid function outputs continuous values of 0-100%. We clarify this point in the revision:

We reconstruct IceNet's output layer by replacing the original softmax with the sigmoid activation function. The sigmoid function outputs continuous values of 0-100%.

**Comment 6**: For Figures 2, 4, 6 and 8, it is slightly confusing to the reader at first how to interpret these results, specifically the Difference column. Perhaps including the caption that red signifies improvement in accuracy and blue signifies a decrease would aid in understanding especially because in Figures 3 and 5 the opposite color scheme is used (red = high error, blue – lower error).

**Response:** Thanks for the comment. We have added the following statement to the captions of Figures 2, 4, 6, and 8: The red signifies a high/improvement in ACC/BACC, and the blue signifies a decrease.