Dear Editor and Reviewer:

Thank you very much for your careful reading, insightful comments and constructive suggestions concerning our manuscript "IceTFT v 1.0.0: Interpretable Long-Term Prediction of Arctic Sea Ice Extent with Deep Learning" (ID: gmd-2022-293), which would greatly help us improve both the content and the presentation of our work. We have carefully considered all comments and are revising the manuscript accordingly. Below all reviewer comments are produced in blue. Our comments are in black, and changes in the manuscript are provided in indented quotes with line numbers.

General comments:
However, the overall structure of the ms is not very reasonable and rigorous. For example, the introduction should not overly emphasize the present study and main point, but focus on reviewing the methods and shortcomings of SIE predictions in the previous studies, then gradually introduce one's own work and emphasize the advantages of the current work.

Thank you for your assistance in improving the manuscript. We re-wrote the introduction to remove the content related to the dynamical models and data assimilation. The current content of the introduction has been revised to introduce the importance of sea ice and the difficulties of SIE prediction, followed by an introduction to the current state of development of SIE prediction models through SIO, and then an introduction to the current machine learning based prediction of SIE based on previous work, summarising the limits of the models are single step prediction, short-term prediction and interpretability of models. Finally, the work and contributions of this paper are presented (pg 1-4, line 19-84).

1. Line 95, the contributions of this paper are not concise enough. For example, (1) is repeated with (2), (3) is a part of (4).

Thank you very much for your comment. We merged contributions #1 and #2, and merged #3 and #4 (pg 3-4, line 73-79):

1) The IceTFT model uses LSTM encoders to summarize past inputs and generate context vectors, so it can directly provide a long-term prediction of SIE for up to 12 months. And it can predict September SIE 9 months in advance, which is longer than other studies with lead time of 1-3 months. IceTFT has the lowest prediction errors for hind-cast experiments from 2019 to 2021 and actual prediction in 2022, which compared with SIO.

2) The IceTFT model is interpretable. It can automatically filter out spuriously correlated variables and adjust the weight of inputs through VSN, reducing noise interference in the input data. At the same time, it can also explore the contribution of different input variables to SIE predictions and reveal the physical mechanisms of sea

ice development.

We have added the reference about the definition of SIE (pg 7, line 132-134):

> SIE is defined as the total area covered by grid cells with SIC > 15%, which is a common metric used in sea ice analysis (Parkinson et al., 1999).

3. The contemporaneous correlation in Table 1 is calculated between the global mean monthly one of eleven variables (e.g., SST) and SIE? And what's the lag-correlation since we are more concerned about the prediction not simulation?

Yes. The contemporaneous correlation is only a preliminary indicator that we use to expand on more input data and is not an important indicator for analysing predictions of SIE. We calculated the autocorrelation of SIE, as shown below:
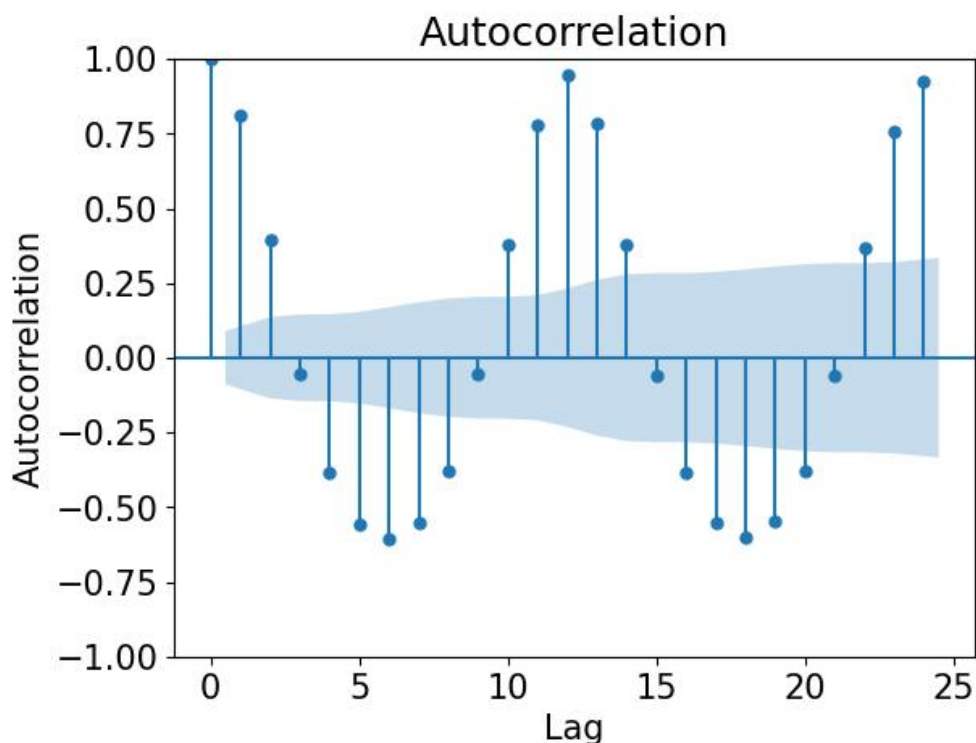
Figure 1. The Autocorrelation of SIE

The x-axis represents the time lag, the y-axis represents the autocorrelation of SIE, and the shaded area is the 95% confidence interval; if the data lies outside the confidence interval, it is autocorrelated. As can be seen from the Fig. 1, the data are all outside the confidence interval, which means that they are all autocorrelated, so that historical data of SIE can be used to predict the SIE. The SIE is highly cyclical, with a 12-month period. It can also be seen from the Fig. 1 that its self-correlation varies with a period of 12. Therefore, with the predicted variable SIE as the primary

predictor, we consider the model with an input step of 12 and do not concern ourselves with the lag-correlation of the other auxiliary predicted data. For the other works with using multivariate to predict sea ice which are presented in the introduction, they also only focus on sea ice when considering the input length.

4. The metrics (MAE, RMSE, RMSD) have units. What's the meaning of the percentages and those in parentheses in Table 3 and Table 4? Please describe the calculation.

The percentages have no special meaning, they are just to show the data more aesthetically. To avoid any misunderstanding, we have reworked it (pg 13).

Table 4. The three metrics (MAE, RMSE, RMSD) among three models on SIE predictions during 2019-2021.

Predictive Year		2019			2020			2021		
Model Name		MAE	RMSE	RMSD	MAE	RMSE	RMSD	MAE	RMSE	RMSD
IceTFT-2018	best	**0.1649**	**0.1942**	**0.4554**	-	-	-	-	-	-
	mean	0.2126	0.2668	0.4756	0.3016	0.3808	0.6182	0.1990	**0.2475**	0.5782
IceTFT-2019	best	-	-	-	**0.2007**	**0.2478**	**0.4890**	-	-	-
	mean	-	-	-	0.2847	0.3747	0.5894	0.2545	0.3345	0.7759
IceTFT-2020	best	-	-	-	-	-	-	**0.1684**	0.2677	**0.6689**
	mean	-	-	-	-	-	-	0.2577	0.3018	0.7071

5. Please give the detailed meaning of the x-axis in Figures 5,6. Time along the x-axis means the started month or target month? How to obtain Error and SIE along the y-axis?

The x-axis is the target month in these two figures.

The line graph represents the observations and SIE predictions, corresponding to the y-axis on the right; the bar graph represents the errors of the observations and the SIE predictions, corresponding to the y-axis on the left.

Now we added in the title of the Figure. 7 (Figure 6. from the previous version of the ms has now become Figure. 7) (pg 14).

> The SIE predictions, observations, and the monthly errors during 2019-2021. The line graph represents the observations and SIE predictions, corresponding to the y-axis on the right; the bar graph represents the errors, corresponding to the y-axis on the left.

6. Many figures and tables are not cited in ms, such as table 4, figure 5 and figure 6.

Thank you for your assistance in improving the manuscript. Although all the figures and tables are cited in the manuscript.

Table 4 shows the three metrics (MAE, RMSE, RMSD) among three models (IceTFT-2018, IceTFT-2019,IceTFT-2020) on SIE predictions during 2019-2021, and it is cited in (pg 13, line 234).

Figure. 6 (Figure 5. from the previous version of the ms has now become Figure. 6) shows the prediction errors of the IceTFT with different input length (6,12,18,24) for 2019, and it is cited in (pg 10, line 210).

Figure.7 (Figure 6. from the previous version of the ms has now become Figure. 7) shows   the SIE predictions, observations, and the monthly errors during 2019-2021, and it is cited in (pg 15, line 251).

7. What's the INITIAL experiment in Table 5? The INITIAL experiment seems to be described near Line 388 but the table 5 is first cited in Line 368. Similar issues exist in other figures.

The INITIAL experiment is the sensitivity experiments with 11 variables. (It has been renamed 11var in the new ms.)

We have restructured the entire article to ensure that the abbreviation is always preceded by the full name of the abbreviation.

8. What's PRATE?

PRATE is precipitation(PRECIP).

We have completed the revision to standardise the abbreviations for precipitation throughout the manuscript to PRECIP.

9. The authors mentioned that VSN in IceTFT can filter the spurious correlations (Line 137). In Section 5.7, the authors only analyze the variable sensitivity. Results in Tables 5 and 6 reflect some common key variables (such as SST and DSWRF), and also demonstrate some individual high-correlated variables (such as PRATE in 2019). From the perspective of statistics, can I consider that the individual high-correlated variables are belong to spurious correlations? Please add some explanations on such phenomenon under different ice cases from the perspective of dynamics.

If a variable makes a large contribution to the forecast, it should have a high sensitivity and can be considered to be highly causal. However, if the variable is highly correlated in Table 1 but has a low sensitivity value in Tables 5 and 6, it can be considered a spuriously correlated variable (such as DLWRF). For PRATE (PRECIP) , it is not considered highly correlated and not very sensitive in 2019-2021. PRATE

(PRECIP) can be considered less causal for SIE. In our analysis, we found that latent heat exchange alters the cloud, allowing for an increase in DLWRF and PRATE (PRECIP), so that PRATE (PRECIP) is the result of other physical factors. We added some explanations (pg 21-22, line 366-377):

> While DLWRF is highly correlated in Table. 1 but has a low sensitivity value in Table. 5, it indicates that this variable is not the cause of the sea ice change, but may be the effect due to other variables. Other studies have shown that latent heat exchange causes more water vapor and clouds to be present in the atmosphere. This enhances the atmospheric greenhouse effect and results in an increased emission of DLWRF. In addition, the increase in water vapor and clouds will lead to more PRECIP. Therefore, there is a correlation between DLWRF and PRECIP, and their sensitivity values change in agreement which both have a higher sensitivity in 2019 and are lower in other years.The positive feedback effect, along with the DLWRF, affects the development of sea ice (Kapsch et al., 2016). Since the machine learning model lacks the partial differential equations of the dynamical model, it cannot simulate the variation of clouds in positive feedback. Therefore, it is difficult to assist in SIE prediction based only on the data trends in DLWRF.

10.From figure 8, the authors compare IceTFT with other models in SIPN. I focus that the others mainly belong to dynamical numerical models and there are no deep learning models. This also means the authors have not compared IceTFT with ordinary deep (machine) learning models. I don't have a clear quantification on how much improvement of TFT and others (such as MLP, LSTM). It is quite a consensus that a deeper (more complex) network is better, but I still want to know if this improvement matches the time and computation comsumption.

We constructed the LSTM model and tuned the model to optimality. And we trained the LSTM model 20 times, taking the best and the mean prediction results for comparison. The results are shown in the table below:

Table 1. The mean results of IceTFT and LSTM model

RMSE	IceTFT	LSTM
2019	0.2668	0.5813
2020	0.3747	0.5715
2021	0.3018	0.6866

Table 2. The best results of IceTFT and LSTM model

RMSE	IceTFT	LSTM
2019	0.1947	0.3864
2020	0.2478	0.4964

2021	0.2677	0.5738

As can be seen from the table, LSTM has a much higher prediction error than IceTFT. By our calculations, the LSTM takes 5 minutes to train 20 times, while the IceTFT takes 5 minutes to train once. But they both take less than 1 minute to test. Although IceTFT takes more time to train, the training time is only in the order of minutes, and small prediction errors can be obtained. Therefore the benefit is worthwhile.

In the study of original TFT (Lim et al. 2019) , TFT has been shown to be applied to time-series prediction of data in various domains, and it can captures long-term features through LSTM encoder and multi-headed attention mechanism. So TFT outperforms other traditional models (such as Recurrent Neural Networks (RNN), Convolutional Transformer (ConvTrans) etc).

In addition, the models in SIPN include not only dynamical models, but also statistical methods, machine learning, etc. The types and the used data of models which are compared in the ms are listed in the Appendix A. Therefore, IceTFT also outperforms other machine learning models in SIPN.

Appendix A: The Data used in SIO Models

Model	Type	Data
NSIDC (Meier)	Statistical	SIE
IceNet1(not include 2019)	Machine Learning	climate simulations (CMIP6), OSI-SAF SIC and ERA5
Sun, Nico	Statistical	SIC, CryoSat-2 SIT
RASM@NPS (Maslowski et al.)	Dynamic Mode	NOAA/NCEP CFSv2, CORE2 reanalysis
NASA GSFC	Statistical	SIE, SIC
UTokyo (Kimura et al.)	Statistical	SIC
Lamont (Yuan and Li)	Statistical	SIC, sea surface temperature (ERSST), surface air temperature,GH300, vector winds at GH300 (NCEP/NCAR reanalysis)
ANSO IAP-LASG	Dynamic Model	wind components (U and V), Temperature (T) in atmosphere and potential temperatur
Climate Prediction Center	Dynamic Model	SIC, Climate Forecast System Reanalysis (CFSR)
CPOM UCL (Gregory et al)	Statistical	SST(ERA5 reanalysis)
CPOM	Statistical	ice area covered by melt-ponds
University of Washington/APL	Dynamic Model	SIC, CryoSat-2 SIT, SST
FIO-ESM (Shu et al.)	Dynamic Model	SST, sea level anomaly(SLA)
Met Office (Blockley et al.)	Dynamic Model	FOAM/NEMOVAR, MO-NWP/4DVar, SIC

Model	Type	Data
PolArctic	Machine Learning	SIE
AWI Consortium (Kauker et al.)	Dynamic Model	SIC,CryoSat-2 SIT, NCEP-CFSR, NCEP-CFSv2
NMEFC of China (Li and Li)	Statistical	SIE, SIC
Wu, Tallapragada and Grumbine	Dynamic Model	NCEP SIC Analysis for the CFSv2, NCEP GFS, GFDL MOM4,Modified GFDL SIS
McGill Team (Brunette et al.)	Statistical	sea level pressure(SLP), area of ice exported through Fram Strait
ARCUS Team (Wiggins et al.)	Dynamic Model	CryoSat-2 SIT, SIC and SST (MERRA-2 atmospheric reanalysis)
ASIC, NIPR	Statistical	SIC, ice thickness, ice age, mean ice divergence
ArCS II Kids	Heuristic	SIC
Cawley, Gavin	Statistical	SIE
CSU-REU21	Statistical	ERA5, Pan-Arctic Ice Ocean Modeling and Assimilation System(PIOMAS)
EMC/NCEP (UFS)	Dynamic Model	NSIDC NASA Team Analysis
GFDL/NOAA (Bushuk et al.)	Dynamic Model	towards 3-D temperature, wind, and humidity data (CFSR), OISST
HEU Group (Zhao, et al.)	Statistical	SIC
Horvath, et al.	Statistical	SIE, ERA5
Kondrashov, Dmitri (UCLA)	Statistical	SIE
KOPRI (Chi et al.)	Machine Learning	SIC
LPHYS2268 - CDDF	Statistical	sea ice volume(SIV), SIT
METNO-SPARSE-ST (Wang et al.)	Statistical	SIE
MetService (Yizhe Zhan)	Statistical	SIE, top-of-atmosphere(TOA), reflected solar radiation (RSR)

11. The authors construct the model based on the original TFT (Lim et al. 2019), please give a detailed parameter configuration, including the hidden layer number and hidden dimensions. Is the current configuration the best?

It is the best configuration. The script code for the call reference is submitted in zenodo.

We add the detailed parameter configuration in (pg 6, line 116-118):

> We have experimentally determined the optimal hyperparameters, the size of hidden layers is 160, the bach size is 128, the number of multi-head self-attentive is 4, the dropout rate is 0.1, the max gradient norm is 0.01 and the learning rate is 0.001.

The IceTFT model has difficulty predicting those extreme years, such as 2020. This predictive limit is described in (pg 13, line 237-242):

> Compared to the results of these two years, the errors of the mean predicted results increased in 2020. This is because there is a second record-low SIE in September 2020. Moreover, due to the predicted period being too long relatively, evaluating the prediction skill of the IceTFT model using MAE as the loss function is difficult. A low MAE does not mean that the model can predict all 12-step with low errors. The IceTFT model focuses on different physical factors during some training, and generates predictions with different trends. The model is hard to predict this minimum value accurately in each training, so the errors of mean prediction are much higher than the best one.

In addition, the SIE has strongly cyclical, the IceTFT with large model errors when the SIE trend is more volatile, i.e. when the slope is larger. In addition, the model has difficulty predicting those extreme years, such as 2020. These predictive limits are described in (pg 15, line 251-260):

> As can be seen, the predictions of multiple training form a predict period in which the vast majority of observations fall within the range. Except for September 2020, the mean predicted results have the same trend as the observations. In terms of the monthly error of the model with different settings, all the experiment runs had high errors in October or November. In addition, they had another high error in July, except for 2019. Due to global warming, it is a challenge to predict SIE in summer. In the melt seasons, which is from June to September, the SIE continued to decline with steep slope. The line passing through the observed value of SIE in June and July has the steepest slope. It demonstrates that the SIE reduced significantly from June to July. Thus, it is difficult to predict the downturn. And as a result, the July prediction is higher than observation with higher error. The SIE archive minimum in September, and sea ice becomes frozen after that time. Similarly, as temperature anomaly or other climate effect, the October or November prediction is on the high side.

The reasons for IceTFT surpassing other models are described in (pg 4, line 86-92):

> Deep learning has good performance in time series prediction, but previous research mostly used CNN, ConvLSTM, which still have high prediction errors. The transformer model makes the attention mechanism fully capture the temporal dependence, and it performs better than the traditional Recurrent Neural Network (RNN) models. Based on the transformer model, temporal fusion transformer (TFT) was proposed for multi-step prediction. TFT not only uses a sequence-to-sequence layer to learn both short-term and long-term temporal

relationships at the local level, but also uses a multi-head attention block to capture long-term dependencies. The TFT has been verified that it has small prediction errors in several areas.The sea ice dataset is a time series with pronounced periodicity, which has a peak and a trough in a yearly cycle, these two peaks are usually critical to the prediction. Therefore, TFT, which can capture long-term temporal features, is suitable for sea ice prediction.

12.Why a higher sensitivity value indicates that the variable makes significant contributions to predictions? I don't understand why the sensitivity can be estimated by adding random noises. To investigate the contribution of different variables to SIE prediction in the model, one can artificially weaken the signal of the concerned variable and investigate the changes in prediction skills.

The basic idea of weakening or enhancing the signals of the variables to investigate the changes in prediction skills is similar. If the variable has a significant effect on prediction, then the prediction error changes as the signal of the variable changes.

$$y^{'} = F(w_1 x_1, \cdots, w_i x_i, \cdots, w_n x_n) \qquad (1)$$

$$y^{'} + \Delta y_i = F(w_1 x_1, \cdots, w_i(x_i + \Delta x_i), \cdots, w_n x_n) \qquad (2)$$

$$\text{Sensitivity}(x_i) = \frac{\text{RMSE}(y, y^{'} + \Delta y_i)}{\text{RMSE}(y, y^{'})} \qquad (3)$$

As in Eq.1. F() represents the deep learning model, x_i represents the input variables, w_i is the weight of the corresponding variable, and $y^{'}$ is the prediction result. For the IceTFT model, VSN automatically adjusts the weights of the input variables during training. If the variable x_i has a significant effect on the prediction, the weight w_i of this variable in the model will be larger. If random noise Δx_i is added to this variable, the change of the prediction result Δy_i will also be larger (Eq.2), increasing the change of RMSE , and therefore the sensitivity (x_i) obtained will be large (Eq.3). However, if the variable x_i has little effect on the prediction, the weight w_i of this variable in the model will be close to 0. When random noise Δx_i is added to x_i, the change of Δy_i is also close to 0. The change in RMSE is not significant, so the obtained sensitivity value is close to 1. Therefore, the higher sensitivity value indicates that the variable makes a significant contribution to predictions.

Kim et al. (2020) added random Gaussian noises to inputs and calculated the change in RMSE to evaluate variable sensitivity. Zhou, L.and Zhang, R (2023) perform sensitivity analysis by assigning a zero value to the temperature anomaly. So

artificially changing the signals of the variables of interest can be used to study changes in predictive skills.

Kim, Y. J. , Kim, H. C. , Han, D. , Lee, S. , & Im, J. . (2020). Prediction of monthly arctic sea ice concentrations using satellite and reanalysis data based on convolutional neural networks. The Cryosphere(3).

Zhou, L., & Zhang, R. (2023). A self-attention–based neural network for three-dimensional multivariate modeling and its skillful ENSO predictions. Science Advances, 9.

Other suggestion:

13.Line 35: "showed" => "shown, "… the errors were …" => "… that the errors were …"

Thanks to your suggestion, the sentence has been amended after our quote rewrite as follows (pg 2, line 32):

> Figure 1 a (b, c) shows that the September SIE prediction errors with a lead time of 3 (2,1) months during 2019 to 2021, which are published in the Sea Ice Outlook (SIO) by Sea Ice Prediction Network (SIPN).

14.Line 40: changed as "there is still a certain gap between these forecasts and observations"

We have fixed it (pg 2, line 36):

> From Fig.1, it can be seen that there is still a certain gap between these predictions and observations.

15.Line 71: "recursive" => "a recursive"

We have fixed it (pg 3, line 44-45):

> Then they used a recursive approach to make the prediction model provide 12-month predictions.

16.Line 80: "lack" => "the lack"

We have fixed it (pg 3, line 58-59):

> Compared to dynamic models, deep learning models are considered a "black box" due to the lack of physical mechanisms.

Thank you for your suggestion, this sentence no longer exists after we rewrote the introduction.

We have corrected all spellings of the word in the ms.

We have fixed it (pg 6, line 120):

> The IceTFT model is designed to use static covariate encoder to integrate static features, and use GRN to generate different context vectors that are linked to the different locations.

We have fixed it (pg 7, line 128):

Each head can learn different temporal features and attend to a common set of inputs.

Thank you for your suggestion, this sentence no longer exists after we rewrote the introduction.

We have fixed it (pg 7, line 137):

> we select a number of variables to support the proposed model for SIE prediction to help it learn more physical mechanisms and improve its prediction skills.

We have fixed it (pg 7, line 153):

> we select these variables: 2m air temperature (AT), DSWRF, DLWRF, and SHUM these variables.

We have fixed it (pg 15, line 256):

In the melt seasons, which is from June to September, the SIE continued to decline with steep slope.

25.Line 295: "As a results" => "As a result"

We have fixed it (pg 15, line 258):

And as a result, the July prediction is higher than observation with higher error.

26.Line 332: "lead time" => "lead times"

We have fixed it (pg 18, line 319):

Figure 10 shows the 2022 SIE predictions of different models of SIO in different lead times.

27.Line 395: delete "has"

We have fixed it (pg 22, line 388):

From Table.6 and Table.7, it can be seen that all six variables had a high sensitivity in the 11var experiment, but the sensitivity changed in the 6var experiment.

28.Line 396: "the more errors" => "more errors"

We have fixed it (pg 22, line 390):

These changes cause more errors in summer(JJAS), autumn(OND), but fewer errors in winter(JFM), as can be seen from the first row of Table.8 (a).

29.Line 403: "on 2021 than 2019" => "on 2021 than on 2019"

We have fixed it (pg 22, line 397):

It demonstrates that the factors affecting the 2019 predictions are similar to those for 2021, and SST and AT have a greater impact on 2021 than on 2019.

30.Line 450: "predict the next year SIE" => "predict the SIE in the next year"

Thank you for your suggestion, this sentence no longer exists after we rewrote the introduction. The similar sentence is as follows (pg 23, line 435):

The IceTFT model can provide the 12-month SIE directly according to the inputs of the last 12 months.

31.Line 451: "the previous year data" => "the previous year's data"

Thank you for your suggestion, this sentence no longer exists after we rewrote the introduction. The similar sentence is as follows (pg 23, line 435):

The IceTFT model can provide the 12-month SIE directly according to the inputs of the last 12 months.

32.Line 456: "IceTFT model" => "the IceTFT model". Missing "the" in many places of the ms.

Thank you for your assistance in improving the manuscript. We checked the grammar of the ms and corrected all the errors.