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.

1. Lines 3-4: In fact, sea-ice melting does not raise the sea level.

Thank you for assisting us in correcting this issue. We re-wrote these sentences(pg 1, line 2-4):

> Due to global warming, Arctic sea ice extent (SIE) is rapidly decreasing each year. According to the International Panel on Climate Change (IPCC) climate model projections, the summer Arctic will be nearly sea-ice free in the 50s of the 21st century, which will have a great impact on global climate change. As a result, accurate predictions of Arctic sea ice are of significant interest.

2. The authors used two words, forecasting/forecast and prediction/predict, in the manuscript. As the timescales are different between forecast and prediction, I recommend that authors use predict/prediction in the manuscript.

We have completed the revision of the entire manuscript and unified it as prediction/predict.

3. Line 13: The authors only gave the prediction results in advance 9 months for 2021, and they should clarify more accurately in the abstract in case of misleading readers. The authors can evaluate more cases for lead times as 9 months.

In the IceTFT, it can generate next 12 months predictions according to last 12 months. For the 12 steps of predictions, the lead time is different for each step. The output of the 1st step is only one month ahead, while the output of the 12nd step is twelve months ahead. To avoid misunderstandings, we re-wrote the abstract to emphasise that the model predicts 12-month SIE directly, and clarify the inputs and outputs of the model(pg 1, line 10):

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

What’s more, we discuss the prediction results of IceTFT from 2019 to 2022. It contains both 3 cases for hind-cast experiments(2019-2021) and 1 case for actual predictions(2022). For the results of hind-cast experiments, it shown in Table 4 (pg 13) , Figure 7 (pg 14), Figure 8 (pg 16) and Figure 9 (pg 17). For the actual predictions experiment, we submitted to SIO the prediction of 2022 September SIE in 2022 June, and the results shown in Figure 10 (pg 19).

4. Line 15: has some physical interpretability -> has a physical interpretability

We have fixed it (pg 1,line 16):

> This confirms that the IceTFT model has a physical interpretability.

5. Line 37: Wei et al. (2021) -> (Wei et al., 2021)

We have fixed it (pg 2, line 34-35):

> And it represents the current predict level and community knowledge of the state and evolution of Arctic sea ice on the sub-seasonal-to-seasonal (S2S) timescale (Wei et al., 2021).

6. The introduction is too long and redundant. For example, in Lines 49-54, it seems that the data assimilation is not close to the manuscript's key point and can brief these to one sentence. Lines 37-38, what's the purpose of "For example, the average SIE…". The authors should reorganize the introduction structure.

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).

7. Lines 94-97: It seems there is a high overlap between contributions #1 and #2. It'd be better to merge them into one.

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.

8. There is a missing part about the description of the data used in the study. It could be added after section 2.

Thank you very much for your comment, although the "Selecting Predictors" section that after Section 2 describes the data used and the reasons for their selection.

9. Figure 2, it’s better if authors list the variables used in the IceTFT framework and give the output clearly (similar to illustration input SIE).

We have modified the IceTFT framework(Figure 2, pg5):

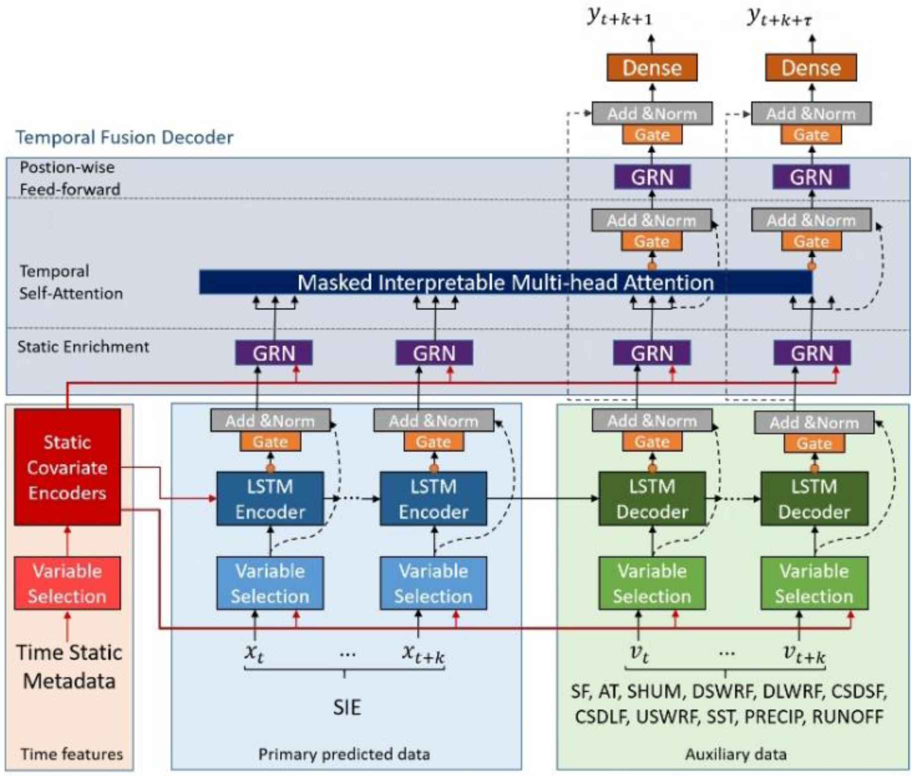


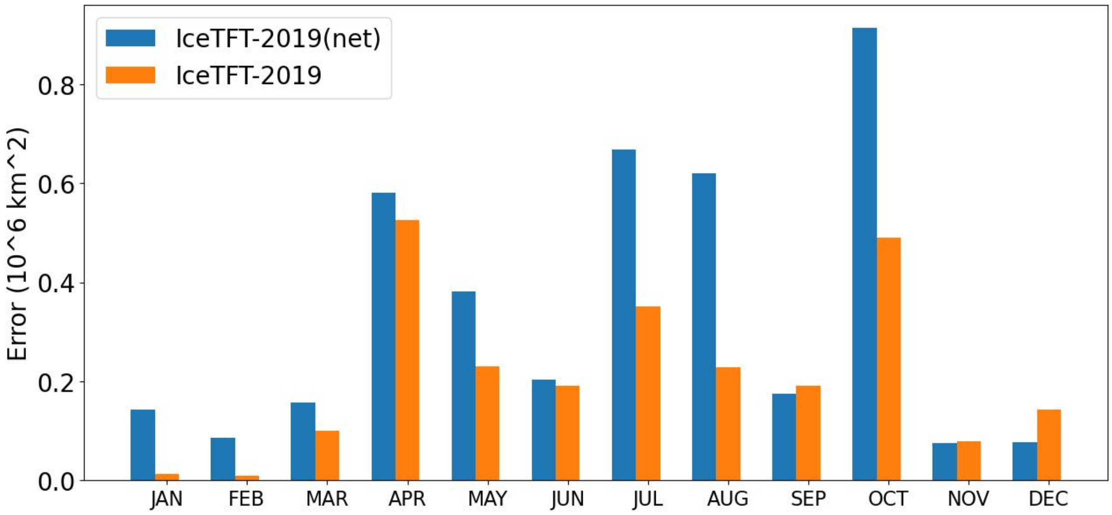
Figure 2. The IceTFT architecture is adapted on the basis of original TFT(Lim et al., 2021). The static time metadata, historical SIE data and other atmospheric and oceanographic variables are all inputs to the IceTFT. The auxiliary data include snowfall (SF), 2m air temperature (AT), 2m surface air specific humidity (SHUM), downward shortwave radiative flux (DSWRF), downward longwave radiation flux (DLWRF), clear sky downward longwave flux (CSDLF), clear sky downward solar flux (CSDSF), upward solar radiation flux (USWRF), sea surface temperature (SST), precipitation (PRECIP) and river runoff (RUNOFF).

10. Line 158: 39.23°-90°N?

We revised it to ‘39.23°N-90°N’ (pg7, line 135).

11. Figure 3: the variables in this figure do not match the variables in the IceTFT framework, such as SW, LW. Meanwhile, according to the authors discussed in section 5.7, I wonder if the results become better using the SW and LW instead of DSWRF, CSDSF, USWRF and DLWRF, CSDLF.

Your comments are very meaningful, and we have added new experiments according to your suggestions. Taking 2019 as an example, we conducted the experiment using the SW, LW data instead of DSWRF, CSDSF, USWRF and DLWRF, CSDLF, and the experiment called IceTFT-2019(net). And we ran the experiments 20 times to get the average predictions for comparison. The results are shown below:



It can be seen that the new experiment IceTFT-2019(net) has a much higher error than the original experiment results in most of the months except December. Since SW/LW are the radiation of DSWRF/DLWRF minus USWRF/ULWRF. Although from the physical mechanism net retains the information characteristics associated with the radiation, for the model this means that DSWRF/DLWRF and USWRF/ULWRF are given fixed weights (1 and -1), and as seen in our experiments, their contributions are different. Thus using DSWRF, CSDSF, USWRF, DLWRF, and CSDLF data would give the model more options to adjust the weights of the input data to improve the predictive performance. Therefore, we still choose the original data (DSWRF, CSDSF, USWRF and DLWRF, CSDLF).

12. The NCEP-NCAR Reanalysis 1 was used in this work. I wonder if the results will change while changing the data to ERA5 or JRA-55. In other words, does the framework depend on the dataset?

Thank you! Indeed this is a good point. The influence of data from different sources on the model performance has be discussed with the JRA-55 data suggested in the comments. We have added new experiments with JRA-55 and addeds a new section of "6.2 Impacts of Datasets on Predictions" (pg17-18, line 301-315):

> To investigate whether the prediction results of IceTFT are affected by the source of input data, we replaced the data from the NCEP-NCAR Reanalysis 1 in Table.4 with JRA55. The same experiments were conducted. Different data sources may be associated with different observation errors, but the physical trends embedded in these data are similar. IceTFT model can automatically adjust the weights of the input data during the training process by adaptively learning the features according to the forecast errors. The label data with different errors can affect the prediction error calculated by the IceTFT model and thus have a large impact on the prediction skill. Theoretically speaking, the prediction skill of the IceTFT model is limited by the source of the label data and does not depend on the source of the input data.

Table 5. The three metrics (MAE, RMSE, RMSD) among three models with reanalysis datasets of JRA-55 on SIE predictions during 2019-2021. Except for SST and SF, other inputs were replaced with JRA-55.

Predictive Year		2019			2020			2021		
Model Name		MAE	RMSE	RMSD	MAE	RMSE	RMSD	MAE	RMSE	RMSD
IceTFT-2018	best	**0.1681**	**0.2214**	**0.4936**	-	-	-	-	-	-
	mean	0.2891	0.3659	0.7165	0.3616	0.4858	0.8166	0.2255	0.2959	0.6131
IceTFT-2019	best	-	-	-	**0.2676**	**0.3360**	**0.6406**	-	-	-
	mean	-	-	-	0.4434	0.6585	1.0836	0.2130	0.2479	0.5458
IceTFT-2020	best	-	-	-	-	-	-	**0.1428**	**0.1801**	**0.4272**
	mean	-	-	-	-	-	-	0.1951	0.2203	0.4966

However, the results are shown in Table 4. It can be seen that the best results of the three models are relative to the original results which are from Table 3, but the mean predictions are higher. This indicates that the models can always get the optimal predictions after several training epochs in the hind-cast experiments and are not limited to the datasets. However, the existence of different observation errors in different datasets makes the bias trends of the predictions different, and therefore makes the mean predictions different. Since the prediction errors using NCEP-NCAR Reanalysis 1 are a little smaller, because in this paper we still use the original dataset for the experimental analysis.

13. Line 223: Evaluation -> Evaluation method

Thank you for your comment, we have modified it to "Evaluation metrics" (pg 9, line 180).

14. Tables 3 and 4, what does the percentage mean? The authors should describe the new statistics variable clearly in the manuscript.

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

15.Line 246: By using the short input length (6 months) leads to worse results. So, what if the input length increases to 18 or 24 months?

Thank you very much for your comment. Considering the difficulty of the 12-step prediction and the fact that the cycle of sea ice includes both melting and freezing processes, we decided to divide the 12-step prediction into two segments of 6 steps each to improve the prediction accuracy. However, actual results show that this assumption is not feasible.

At your suggestion, we conducted experiments with different input steps to discuss the model prediction techniques in terms of the input steps of the model. We re-wrote the section 5.2 “The Input Length” (pg10-11, line209-219):

> To investigate the effect of input length on the prediction skill, we chose to set up four sets of comparison experiments with input length of 6,12,18 and 24. Using 2019 prediction as an example, the results of the monthly errors are shown in Fig.6. The results of 2022-2021 are similar and we omit to show them. As a whole, the prediction errors for the models with the input lengths of 6 and 24 are significantly higher than the results for models with other lengths. Probably because the time window of 6 is too short to include both the March maximum and the September minimum in each epoch. This may affect the model learning for the features of the extremes, increasing the inaccuracy of the extremes. However, if the input lengths are too long, the correlation between the recent historical SIE sequence and the future SIE sequence is weakened, increasing the prediction error. In addition, the errors of a model with 18-month are comparable to that with the 12-month, but for the difficult prediction of 2019, i.e., October, which has a large slope, the error of a model with 18-month is significantly higher than that with the 12-month. Therefore, for the monthly prediction of SIE, a reasonable choice for the input length is 12-month, it probably is because the period of SIE is 12-month.

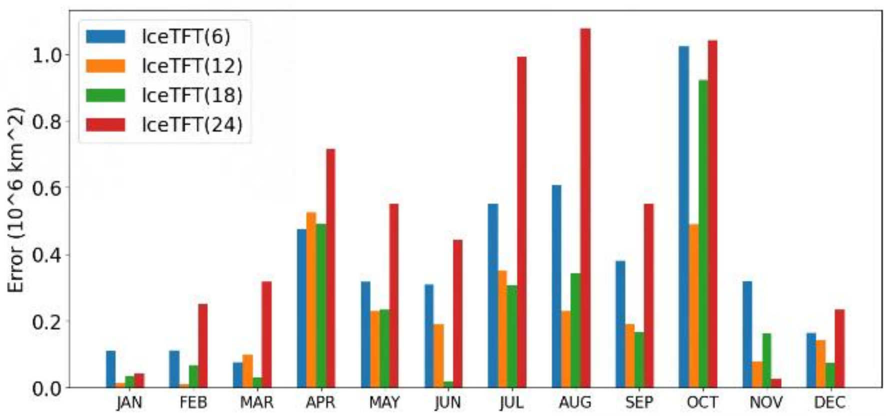


Figure 6. The errors of the IceTFT with different input length for 2019

16.From figure 6, it can be seen that the biases are much larger in Sep than in winter or other seasons.

Figure. 6 (now it is Figure. 7) contains two vertical coordinates, but we do not explain them in detail. Now we added in the title of the 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.

We have discussed the biases, and the September errors are relatively small. The predictive months with large prediction errors are mainly in July, October and November(pg 15, line 251-260):

> From the bar graphs in Fig.7, there is a clear trend of predictions for different years, and it also shows the monthly errors. 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.

So, it's better that the authors can evaluate the IceTFT model's ability in different seasons, which may be more helpful for using the IceTFT model and understanding the sea-ice prediction ability. In fact, the prediction ability in summer (JJAS) is also more important than in other seasons.

We analyzed the predictive limits of IceTFT in SIE prediction in different months.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. These predictive limits are described in (pg 15, line 255-260):

> 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.

We evaluate the IceTFT model's ability in different seasons and explore the potential causes for the inaccuracy between the SIE observations and the predictions according to the RMSD between the detrended quarterly SIE observations and the predictions for the 2019–2021 period (pg 15-16, line 268-283):

> To further explore the potential causes for the inaccuracy between the SIE observations and the predictions, we calculated the RMSD between the detrended quarterly SIE observations and the predictions for the 2019−2021 period. The results are shown in Fig. 8. The RMSD ranges from 0.076 to 0.918 million km2 in Fig.8 (a), and the findings from the three years show a wide spread in RMSD on quarter. Figure 8 (b) displays a histogram of the temporal variation of squared RMSD, consisting of “bias” and “variance” according to Eq. (4). It can be seen that there is a very large variance in the spring (JFM) of 2020 and 2021, which is responsible for the high RMSD in this season. The correlation coefficients in Fig.8 (c) also display an obvious reduction in spring 2020, which is consistent with the variance variations in Fig.8 (b). This result indicates that the significant lower correlation coefficients are partially responsible for the RMSD peak. Moreover, except for a few months, the magnitude of the bias is substantially larger than the variation in Fig.8 (b), indicating that the change in bias is the main factor for the increase in RMSD. Figure 8 (d) shows the standard deviations of the predictions of IceTFT model and observations, and the annual standard deviation represents the amplitude of the seasonal cycle of SIE. The results show that the difference between these two standard deviations is obviously increasing, which contributes to the lager increase in bias over the same period. Furthermore, this is consistent with the finding in Fig. 7. The IceTFT with large model errors when the SIE trend is more volatile, i.e. when the slope is larger, such as in July and October. The biases between predictions and observations are larger for the season containing these two months. This suggests that IceTFT does not fully capture the signals from the historical data and does not reflect the seasonal variability in the SIE. Thus, we can improve the predictive model by focusing on the seasonal variability in the predictions to reduce the RMSD.

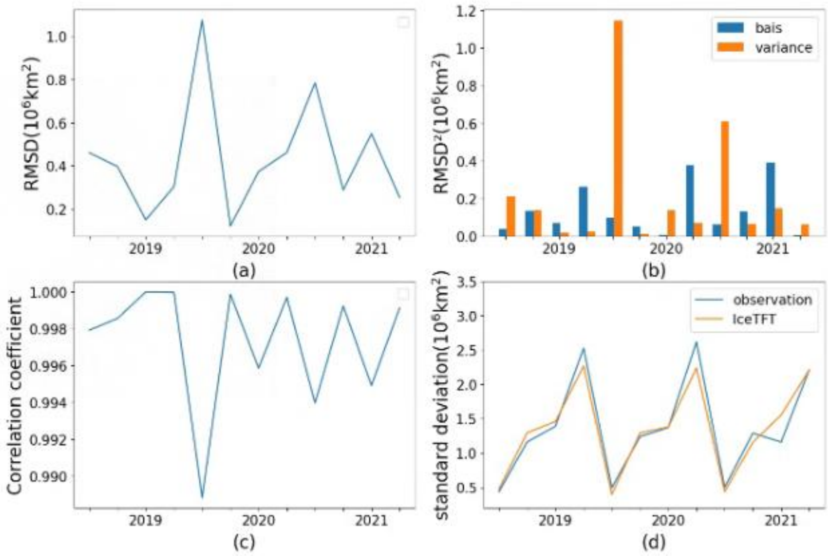


Figure 8. Time series of the RMSD between the detrended quarterly SIE on the IceTFT-model over the period 2019-2021: (a) RMSD; (b) squared RMSD (histogram), consisting of "bias" and "variance"; (c) correlation coefficient between predictions and observations; (d) standard deviation of predictions (orange line) and observations (blue line).

In addition, We have done interpretable analysis for seasonal predictions in Sect.8.3 "Analysis of the physical mechanisms on the seasons", mainly for summer(JJAS) and winter(JFM). Some of the conclusions are as follows (pg 22-23,line 389-433):

> Consequently, during the melting season, a relatively small area of sea ice cover exposes a large area of sea surface, and warming seawater affects sea ice melt. Since our model cannot simulate the process of radiation absorption by the ocean, SST can provide the IceTFT model with a direct factor affecting sea ice melt. However, for the freezing season, when the sea ice cover is large and the exposed sea surface area is small, the effect of SST on sea ice melt is relatively small. Rather, heat fluxes and warming air temperatures from water vapor, cloud cover and radiation mechanisms have a greater effect on sea ice melt (Kapsch et al., 2013; Boisvert and Stroeve, 2015). Thus validating the conclusions of our experiments that SST is an important factor influencing prediction from August to October, while radiation-related variables and AT are from January to May.

17.Figure 8: as we know, SIO is the prediction results from June, July, and August. The authors should clarify what kind of prediction data of SIO used in this figure.

There are more models in the SIO, and we have collected all the models which in the manuscript, recording the types of models and the data they use, listed in Appendix A (pg 25-27):

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)
NCEP-EMC (Wu et al.)	Dynamic Model	NCEP SIC Analysis for the CFSv2
NCAR/CU (Kay/Bailey/Holland)	Heuristic	Mitch Bushuk GFDL for a synthesis project)
NSIDC Hivemind	Heuristic	SIE
Simmons, Charles	Statistical	Moana Loa monthly CO2 concentrations, Northern Hemisphere snow area, SIC
Slater-Barrett (NSIDC)	Statistical	SIC
SYSU/SML-KNN	Machine Learning	SIE, SIC
SYSU/SML-MLM	Statistical	SIC, SST, surface air temperature (SAT), surface net radiation flux (NR)
UPenn-UQAM Group	Statistical	SIE, SIC
UKMO-OIT	Heuristic	-
UQAM (VARCTIC)	Statistical	SIC, SIV