Reply to Reviewer 1- tc-2021-282 Asadi et al. 2021 - Probabilistic Gridded Seasonal Sea Ice Presence Forecasting using Sequence to Sequence Learning

We sincerely thank the reviewer for the thorough review and excellent comments.
Reviewer comments are shown in black, our responses are shown in blue.

This manuscript presents an innovative forecast tool for sea ice conditions (presence) based on a machine learning approach using sequence to sequence learning for both short term (7 days) and long term (up to 90 days) predictions.
The presented tool is, without a doubt, something that is of interest to the sea ice expert community and is based on novel methods of machine learning that make analysis of massive information datasets possible nowadays.

Even though the pertinence of the presented tool, several improvements must be done on the manuscript. The research design in itself has to be described into more details, especially by providing a more complete description of the different tests and protocols followed in the experiments and model calibration part.
In addition, many paragraphs, especially in the methodology part, could be supported by figures and schematic representations, for example, the ML design and architecture.

We have included preliminary figures (shown later in this response) describing the training and testing sequence (model calibration) and also the ML architecture (encoder and decoder).

Also, as the Hudson Bay region is highly documented and studied, the results obtained by your approach could be compared to data provided in the sea ice atlas from the Canadian Ice Service or results from other probabilistic/modelling approaches applied on the Hudson Bay (Saucier et al. 2004, Hochheim and Barber 2014, Kowal et al. 2017, Gignac et al. 2018, Dirkson et al. 2021.). Even though the comparison may not be quantitative, a qualitative assessment, outlining the differences between the approaches and the strategic advantages you provide using ML would be relevant.

Thank you for this comment. Gignac et al (2019) and Dirkson et al (2019) developed methods for probabilistic forecasting of sea ice concentration based on fitting probability distribution functions (PDFs) to historical passive microwave sea ice concentration data. Gignac et al (2019) used a beta PDF. They found their approach was able to capture freeze-up and break-up within one or two weeks of dates provided by the Canadian Ice Service (CIS) ice atlas, with the exception of Sanirajak (Hall Beach), where there is a polynya. Dirkson et al (2019) developed a related approach but used zero and one inflated beta distribution. This distribution allows the endpoints of their probability distribution to be captured differently than the interior points. Their PDF is fit to output from a prognostic modelling system, CanSIPS, (Canadian Seasonal to Interannual Prediction System), which consists of two coupled atmosphere-ice-ocean models. A novel bias correction approach is applied to their predictions and CanSIPS output before comparison with HadISST2 sea ice and surface temperature dataset. Their predictions show skill in Hudson Bay for forecasts initialized in May and June for 1-2 months (their Fig 10). Building on this, a related approach was used in Dirkson et al (2021), where a PDF was fit to freeze-up and break-up dates. In that case, before bias

correction the model (CanSIPS) predictions of freeze up and break up in Hudson Bay were biased by 4 weeks, and 1-2 weeks respectively, where the data used for comparison is passive microwave sea ice concentration. The proposed bias correction methods (Dirkson et al. 2021) lead to improved prediction of these events in Hudson Bay for lead times of 1-6 months, depending on the location. For both of these studies the horizontal grid resolution was 100 km and time scales were monthly.

We also found two other relevant studies, those by Bushuk et al (2017) and Sigmond et al. (2016). Bushuk et al. (2017) use a fully coupled atmosphere-ice-ocean-land model with horizontal grid resolution for the ice-ocean component of 1 degree. They compute the anomaly correlation coefficient (ACC) of detrended sea ice extent from their model with that from passive microwave data and compare with a model of anomaly persistence. Their model shows skill relative to persistence in Hudson Bay, in particular in summer months, with lead times of 3-8 months. Sigmond et al. (2016) use CanSIPS. They examine anomaly correlation coefficients between sea ice advance and retreat dates computed from the model and from passive microwave sea ice concentration data. The grid resolution is relatively coarse (around 100 km), possibly because it is an ensemble forecasting system, which is computationally demanding. They find their model has skill (defined by statistically significant ACC) at lead times of 2-6 months (average 5) for sea ice advance (freeze-up) and 2-3 months (average 3) for sea ice retreat (break-up) (not detrended, results for Hudson Bay). We have included some comments in the revised manuscript to reflect these related studies.

It should be noted while our approach is similar those by Gignac et al. (2019) and Dirkson et al. (2019, 2021), in that it does not use dynamical sea ice models to produce a prediction, it is different in that it uses a sequence-to-sequence learning approach that is capable of producing forecasts. The present configuration is to used for 90-day forecasts, although it is evaluated in the submitted manuscript in hindcast mode to demonstrate the concept.

Overall, the presented method and tools appear as scientifically sound and clear. They should, however, be described more carefully and more examples of applications of the model shall be presented to the readers.

It is a work of great interest and I hope my comments will guide and help you in improving your manuscript.

Thank you very much. Your comments are very helpful.

Major comments

1. As aforementioned, *comparisons have to be made and applications examples provided*. Especially in areas of high variability or in the presence of particular entities such as polynyas or narrower Bays (Frobisher Bay or Hall Beach polynya, for example).

In the revised manuscript we have included a comparison of freeze-up and break-up dates with ice charts for the ports identified in the manuscript (Churchill, Inukjuak and Quataq) as well as; Sanirajak (formerly known as Hall Beach), Kivalliq and Kinngait, where there are polynyas. We have investigated using daily ice charts of the Canadian Ice Service for this purpose. These are sea ice

concentration and ice type analyses derived through manual inspection of SAR data and other data available, with valid time of 18 UTC each day. However, these charts are not available on a daily basis through the entire ice season in the Hudson Bay region. Hence, our analysis is based on regional ice charts, which are similar to daily ice charts, but bring in data over a week and have less fine spatial detail. We also considered using the CIS ice atlas for this comparison (as was used in Gignac et al) but the CIS ice atlas covers an earlier period. Note that the polynya toward the outlet of Frobisher bay is not in our study region. The comparison with the ice charts is in Section 7.2.3 of the revised manuscript.

2. *Limitations of the approach shall be discussed*. The model is providing forecasts on a ~31 km grid. How does this affect the usage capabilities for the principal expected users (the mariners)? This should definitely be discussed more in depth.

The resolution is similar to that used in other studies on seasonal forecasting that have been developed with mariners in mind. For example, passive microwave data were used for development of the probabilistic approach of Gignac et al. (2019) and for validation of subseasonal-to-seasonal (S2S) time scales (Zampieri et al. 2018). While passive microwave sea ice concentration data are often gridded to 25 km, the spatial resolutions of the brightness temperature data used to generate the sea ice concentration are typically coarser. The 19.35 GHz channel on the SSMI and SSMI/S sensors (often used to produce sea ice concentration observations) has an instrument field of view of approximately 45 km x 69 km (https://www.remss.com/missions/ssmi). For studies that carry out seasonal forecasting using a dynamic ice-ocean model (or similar) where a sea ice state vector is predicted as a function of time, our resolution is similar to that used in other approaches to be used to support (in part) marine transportation (Sigmond et al., 2016, Askenov et al 2017). Sigmond et al. (2016) assess seasonal forecasts of sea ice advance and retreat at a spatial resolution of 100 km, while Askenov et al. (2017) examine navigation routes using a ¼ degree grid (nominally 28km, 9-14 km in Arctic) coupled sea-ice ocean model. For the latter study, their focus is on the Northern Sea route, and their study region does not include Hudson Bay or Hudson Strait. Nevertheless, the study by Askenov et al. (2017) study highlights potential future changes in Arctic ice cover that will lead increased navigability, but with significant risk

For trend analysis in the Hudson Bay region, which also supports planning activities, Hoccheim and Barber (2014) use passive microwave data (gridded to 25 km) to look at sea ice trends in the Hudson Bay system, while Andrews et al. 2018, use a combination of passive microwave data for offshore trends in ice cover, and regional ice charts from the Canadian Ice Service for nearshore trends.

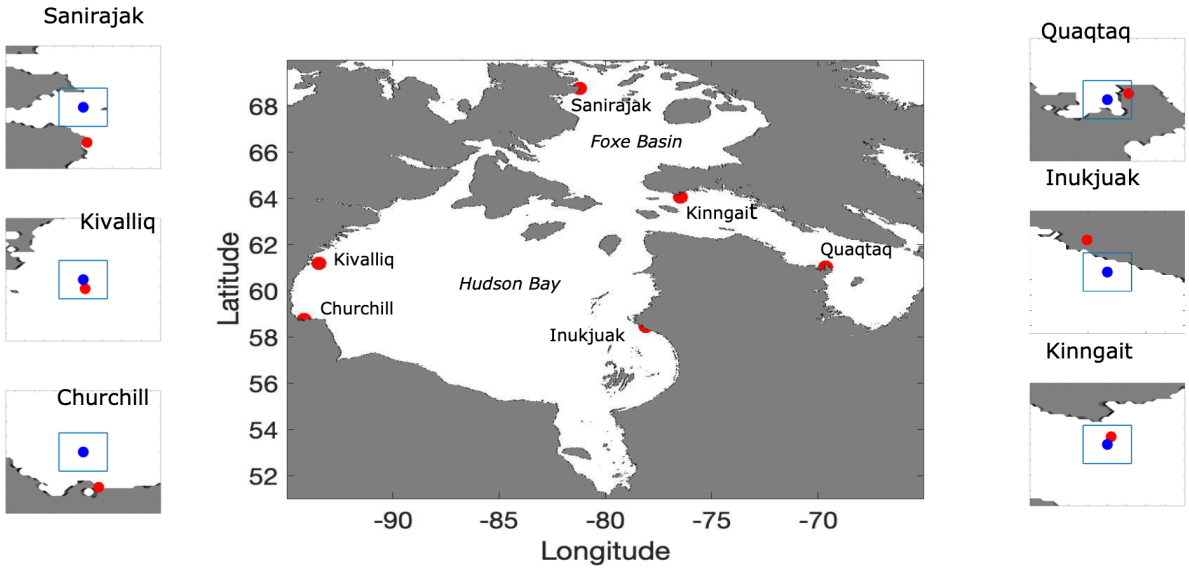
We have added some of these comments to the Discussion Section of the re-submitted manuscript.

3. *Sensibility to the input "sea ice normal condition" wasn't discussed*, when speaking about the augmented model. *Were variable time spans tested?* If so, did they generate similar forecasts? If not, how would you explain this situation? In other words, a certain "sensibility analysis" would be convincing about the model capabilities.

In the version submitted, the augmented model secondary input is based on climate normal computed as the average 2m air temperature (t2m), and average 10m wind components (u10 and v10), where the average is calculated from the first year (1985) until the last year of training (2015). We compared this method of computing the climate normal (CN1) with one that is based on the 10 years previous to the validation year (CN2). There was very little difference in the model predictions when these two different augmented models are used based on heat maps (of the type shown in Fig 1 in the original submitted manuscript). This indicates the patterns our model is learning do not vary significantly over the 1985-2015 period in comparison to a decadal time scale. Examining the results on a year to year basis showed that when the shorter time period is used for climate normal the year-to-year variability has a more significant impact on model performance, indicating climate normal is given less weight, or is more similar to the other model inputs.

4. I strongly suggest that you *add a map of your validation sites* and provide a short description of each. For example, Quaqtaq is located in a bay, narrower than 31km. How this does affect the results?

We have put together a map of the port locations, including one requested at Sanirajak (formerly known as Hall Beach). The map of the study region is shown below with the port locations shown in red. The Insets show the port location (red) and the nearest point on the model grid (blue) that is outside of the land boundary (where landmask from ERA5 is less than 0.6), in addition to a bounding box that approximates a grid cell. The model grid point near Quaqtaq is located (correctly) in the water region because the landmask from ERA5 has a low value in that region due to the low elevation.



Also, *why were these 3 sites chosen*?

The sites were chosen because they represent locations with significantly different sea ice conditions. Churchill and Inukjuak are located on the east and west coasts of Hudson Bay, with Churchill being a major port as part of the potential Arctic Bridge shipping route. The east coast is

significantly impacted by influx of freshwater inflow from rivers draining into Hudson Bay, while the west coast region is impacted by northwesterly winds (there is a latent heat polynya, the Kivalliq polynya, that runs along the northwest shore of Hudson Bay). There are additionally east-west asymmetries in Hudson Bay in terms of ice thickness and sea surface temperature (Saucier, 2014), with counter-clockwise ocean currents leading to thicker ice covers along the eastern shore of the Bay. Quataq is located in Hudson strait, where wind and air temperature patterns are different from those in Hudson Bay, and pressured ice is common in the consolidated ice season. We will be bringing in Sanirajak and a location in the central part of the Bay (away from the coast) in the revised version.

Have you found any irregularities in the ERA-5 sea ice concentration dataset you used to calibrate your model?

We found there were irregularities with the ERA landmask file and the sea ice concentration. There were some locations indicated as land in the landmask file that had a non-zero sea ice concentration value. At these locations the sea ice concentration was set to zero. There were also some locations indicated as non-land in the landmask file that had a zero ice concentration, even when the ice concentration should be non-zero based on the atmospheric conditions and seasons. At these locations (indicated by landmask less than or equal 0.6), the sea ice concentration was set to the non-land average of neighboring pixels.

5. Nowhere in the manuscript have I found a *justification on why it is the presence of ice that is modeled and not the concentration*. That should be discussed since the OSI-SAF OSI409 (which are SIC) data are ingested into ERA-5

The approach used here predicts a grid of (uncalibrated) sea ice probabilities. For seasonal forecasting, probabilistic information is often desired (Gignac et al. 2019), in particular regarding freeze-up and break up (Wagner et. al., 2020). While the model does ingest sea ice concentration, this is combined with other environmental variables to produce a binary probability of ice vs water at the grid cell, or ice presence. This output is similar to probabilistic approaches by Gignac et al. (2019) and Dirkson et al (2019) where their model estimates the probability of sea ice concentration exceeding a certain threshold (e.g., 15%).

Specific comments

Line 10 : Define “high spatial resolution” for the reader. Depending on your field, it differs.

We were referring to the spatial resolution of 5-10 km. For example, the high-resolution ocean and sea ice forecasting system for the Arctic and North Atlantic oceans (Dupont et al., 2015).

Line 46 : Define sea ice presence (SIC > 15%).

We noticed that the term “sea ice presence” was first used at the beginning of the introduction, and have added the information there (sea ice concentration greater than 15%).

Line 51 : This information should be provided way before, otherwise, some will think, as I did, that you use remote sensing data.

The use of ERA5 has been added to the last paragraph of the introduction. The abstract has also been changed to explicitly mention this “Given the recent observations of the declining trend of Arctic

sea ice extent over the past decades, seasonal forecasts are often desired. In this study machine learning (ML) approaches are deployed to provide accurate seasonal forecasts based on ERA5 data as input"

Line 57 : Why not starting in 1979 ?

Initially the study was started with a different data set (other than ER5). These data started in 1985, hence that was used in this study as well because it provides a sufficient time series for training and testing.

Line 67 : Remove last "and". Corrected

Line 87 : A schematic representation of the encoder and decoder parts would be useful.

Below is a figure of the architecture which shows the encoder and decoder parts in more detail. The upper panel shows the overall architecture(described on lines 98-105 of the submitted manuscript, modified here).

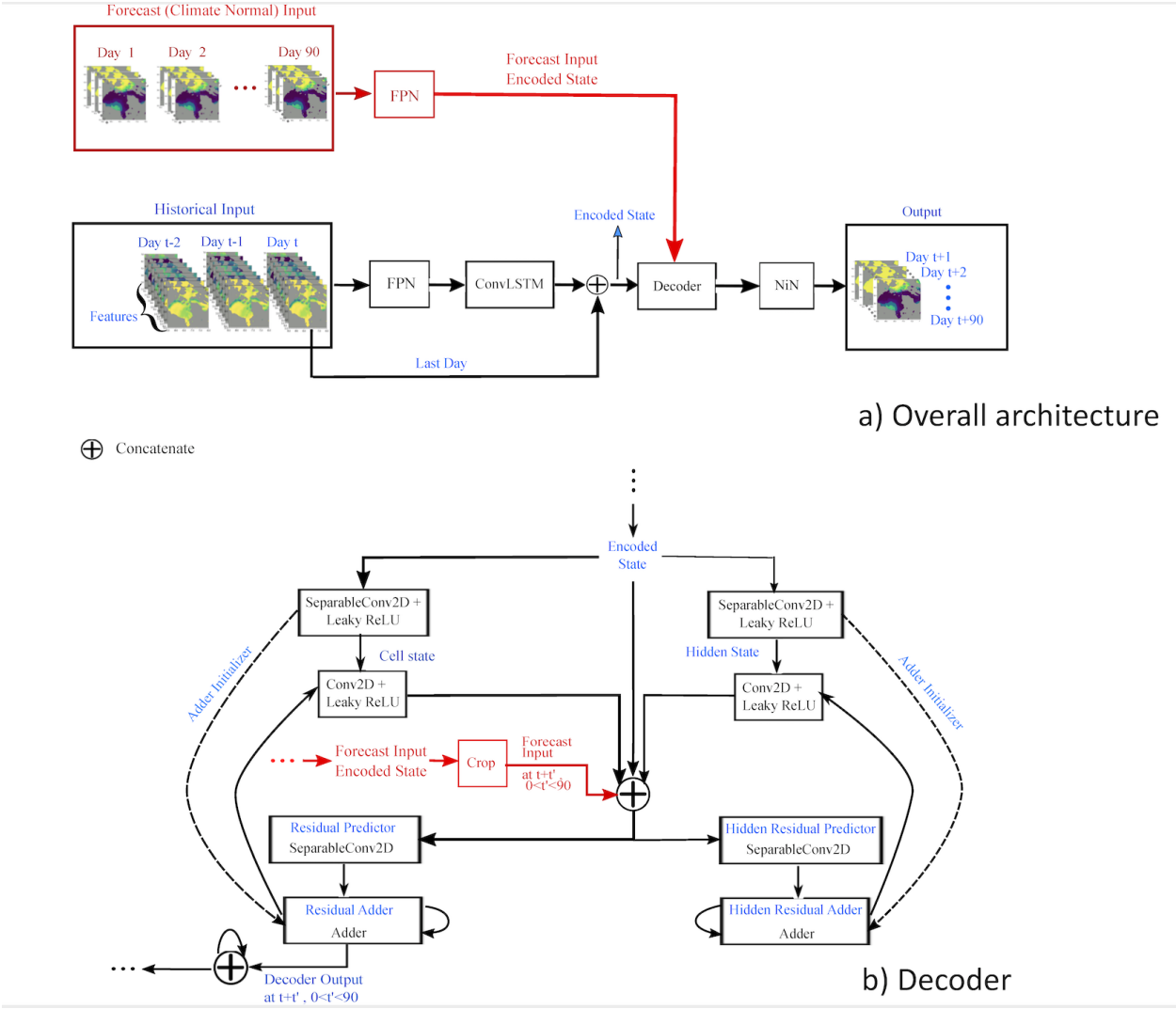
The overall architecture is shown in the figure below (panel a). The encoder starts by passing each daily sample through a feature pyramid network (Lin et al., 2017) so as to detect environmental patterns at both the local and large scales. Next, the sequence of feature grids extracted from the feature pyramid network are further processed through a convolutional LSTM layer (ConvLSTM) (Hochreiter and Schmidhuber, 1997; Xingjian et al., 2015), returning the last output state. This layer learns a single grid representation of the time series that also preserves spatial locality. Finally, the most recent day of historic input data is concatenated with the ConvLSTM output. The encoder provides as output a single raster with the same height and width as the stack of raster data input to the network, but with a higher number of channels such as to represent the fully encoded system state. The final encoded state is then fed to a custom recurrent neural network (RNN) decoder that extrapolates the state across the specified number of time-steps. It takes as input the encoded state with multiple channels and as output produces a state with the same height and width as the input over the desired number of time-steps in the forecast (here 90 days). Finally, a time-distributed network-in-network (Lin et al., 2013) structure is employed to apply a 1D convolution on each time-step prediction to keep the spatial grid size the same but reduce the number of channels to one, representing the daily probabilities of ice presence over the forecast period (e.g. up to 90 days).

The lower panel shows the decoder (described on lines 106-113 of the submitted manuscript, modified here)

The custom RNN decoder, as is common of many RNN layers, maintains both a cell state and a hidden state (Yu et al.,2019). First, the initial cell state and hidden state are initialized with the input encoded state. Then, at each time-step and for each of the states, the network predicts the difference, or residual, from the previous state to generate the updated states using 2D depthwise separable convolutions (Howard et al., 2017). Depthwise separable convolutions are chosen to preserve the time dimension of the input, which the convolution operates over the two

spatial dimensions. The output of the decoder section is the concatenation of the cell states from each time-step (unrolling of the learned time sequence).

The red portion shown in the figure below corresponds to the additional components required for the Augmented model (described on lines 115-122 of the submitted manuscript)



.

Caption: Figure showing (a) the overall network architecture and b) the custom decoder. The red portion refers to the additional components required for the augmented model. The dashed arrows show a process carried out only once (the initialization of the adder). FPN refers to the feature pyramid network, ConvLSTM, the convolutional long short-term memory network, NiN is the network in network module.

Line 92 : What time bin(-s) were used as input (12:00, 00:00, a daily average ?)

Noon samples were used. This information has been added to the last paragraph of the data section.

Line 112 : How does extending to a longer input affects the forecast quality ?

We tested extending to a 5 day input but did not see any improvement in forecast quality. With this longer input the quantity of data to be processed is greater than that for 3 days, which increases the computational expense and data storage requirements. Hence we did not continue with this, or longer inputs. However, we recognize that in a different domain where other processes are important a longer input may show improvements.

Line 126 : Can you describe these “extensive experimentations” ?

Initially we had used a leave-one-out approach for training and testing of the model. For example, given 30 years of data, train with 29 and test with one, and repeat over all 30 years. However, this results in the use of future data for training, which is not desirable for a forecasting approach. Hence, we moved to the current approach of training initially using ten years, updating the model weights for future time periods. We tested different training periods (10 vs 20) and also different numbers of months to include in training our monthly models. The current configuration led to the best results.

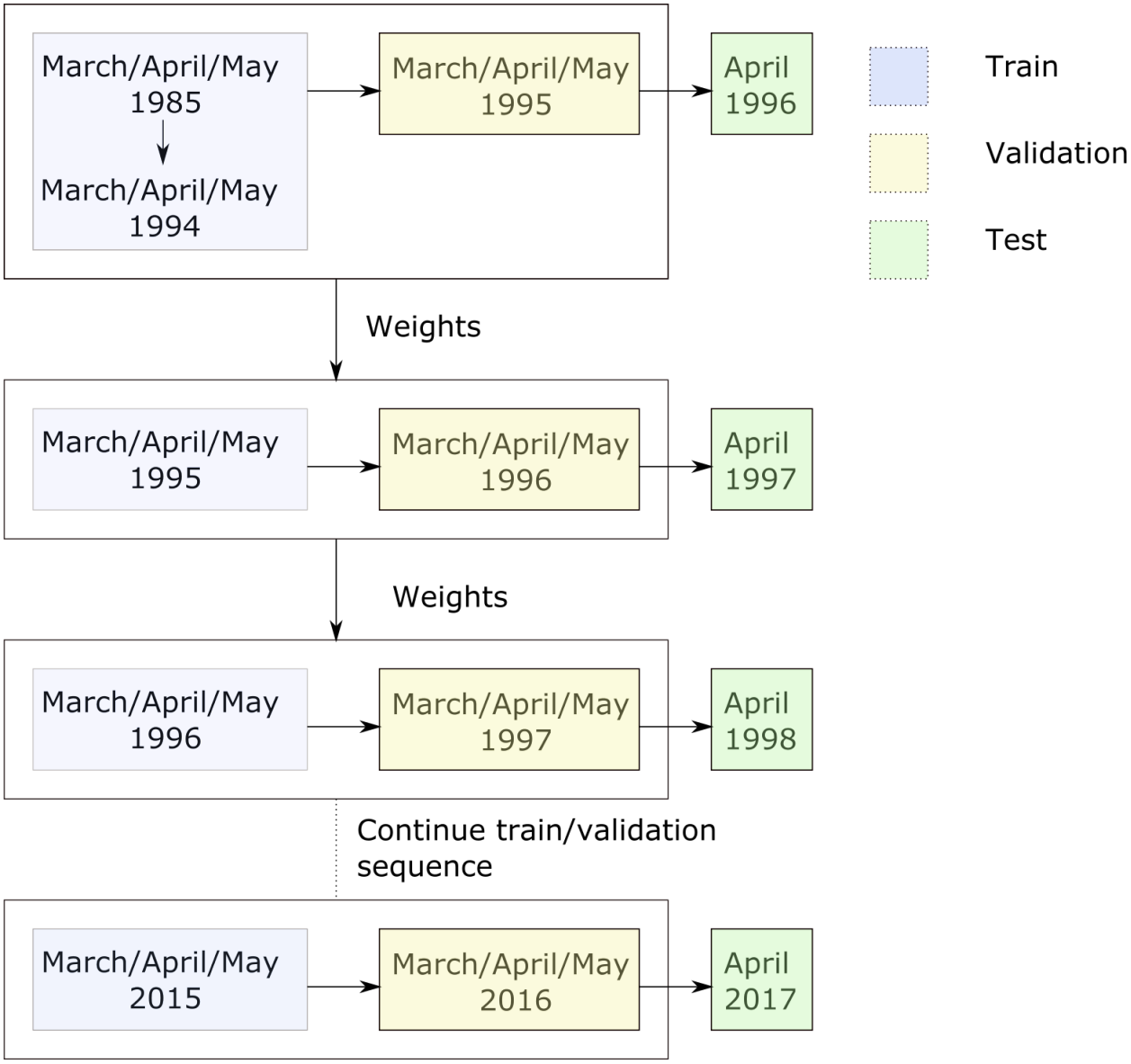
Line 129 : “Chosen to be 10 years”. Why is it so ?

The choice of 10 years is a compromise between having enough data to provide the model with representative conditions from which it can learn, and not having the approach become too data-heavy. Because the model already performs reasonably well with this training approach, it could be used alternatively with other data sets for which a shorter time series is available (eg. an AMSRE/AMSR2 unified data set available from 2002, which would have higher spatial resolution, although also different variables).

Line 135 : This processing logic should definitely be represented in a figure.

Good point. We agree and have prepared a figure (below with text from manuscript lines 127-134, modified here)

For each month of a year a separate model is trained on data from the given month as well as the preceding and following month. For example, the ’April model’ is trained using data from March 1 to May 31. This monthly model is initially trained on data from a fixed number of years, chosen to be 10 years. After this initial experiment, to predict each following test year i, using a rolling forecast prediction, the model from year i−1 is retrained with data from year i−2 and also, data from year i−1 is used as validation for early stopping criteria and to evaluate the training performance. For example, if the initial model is trained on 10 years, data from year 11 is used as validation and first predictions are launched at year 12. The model for year 11 is initialized with weights from the 10-year model and retrained with data from year 10, validated on year 11 and predicts year 12. The model for year 12 is then initialized with weights from the year 11 model, retrained with data from year 11 and validated on year 12 to predict year 13. This process is used to produce forecasts of sea ice presence for years 1996 to 2017.



Line 166 – 167 : The end of the sentence doesn't make sense. Consider reformulating.

This has been changed to "Using additional climate variables for the input of the Augmented model is shown to be beneficial (Fig 1d,e,f). In the periods where the Basic model is worse than climate normal, the Augmented model has better accuracy, for example for the April model at lead days 60-80, and the August model, lead days 75-90..

Line 181 : It seems counterintuitive. Can you explain why ? The reason for the better Brier score of the Basic model in comparison to augmented here may be because the September model uses training data over August/September and October. We hypothesize that the trend over this period may be less representative of more recent ice conditions than breakup, which may make the additional data used in the Augmented model un-helpful.

Figure 1 : Y-Axes for subfigures d-e-f should be Accuracy differences or ΔAccuracy.

This has been changed.

Line 220 : What would you link the lower accuracy in “central region” ? Is it the higher variability of the freeze-up pattern or to climate variables that are, given the distance to stations, less reliable in such areas ?

We assume the reviewer is referring to the lower performance of Basic model in the central regions relative to the coast in December for a 30 lead day forecast. In this case the degradation was because freeze-up was too late in the model in the central regions. While the climate variables could be less accurate due to their distance to station data (assuming station data are assimilated in a reanalysis and are accurate) our experiments are not set up to evaluate this because we are using ERA5 as our “observations” for comparison.

Lines 236, 246 & 249 + Figures 8 & 9 : From what I know, it should be written Quaqtaq, not Quataq (https://www.makivik.org/quaqtaq/).

Thank you very much. This has been corrected.

References

Andrews, J., Babb, D. and Barber, D.G. (2018), “Climate Change and Sea Ice: Shipping in Hudson Bay, Hudson Strait and Foxe Basin (1980-2016)”, Elementa, 6, DOI: 10.1525/elementa.281.

Askenov Y., Popova E.E., Yool, A., Nurser, A.J., Williams, T.D., Bertino, L. and Bergh, J. (2017), On the future navigability of Arctic sea routes: High-resolution projections of the Arctic Ocean and sea ice, Marine Policy, 75, 300-317.

Bushuk, M., Msadek, R. Winton, M. Vecchi, G. A., Gudgel, R., Rosati, A. and Yang, X., (2017), “Skillful regional prediction of Arctic sea ice on seasonal time scales”, Geophysical Research Letters, 44, doi: 10.1002/2017GL073155.

Dirkson, A., Merryfield, W.J., and Monahan, A.H., (2019) “Calibrated probabilistic forecasts of Arctic sea ice concentration”, Journal of Climate, 32, 1251-1271.

Dirkson, A., et al. (2021) "Development and Calibration of Seasonal Probabilistic Forecasts of Ice-free Dates and Freeze-up Dates." Weather and Forecasting 36.1: 301-324.

Dupont, F., Higginson, S., Bourdallé-Badie, R., Lu, Y., Roy, F., Smith, G. C., Lemieux, J.-F., Garric, G., and Davidson, F.: A high-resolution ocean and sea-ice modelling system for the Arctic and North Atlantic oceans, Geosci. Model Dev., 8, 1577–1594, https://doi.org/10.5194/gmd-8-1577-2015, 2015.

Gignac, C., Bernier, M., & Chokmani, K. (2019). IcePAC–a probabilistic tool to study sea ice spatio-temporal dynamics: application to the Hudson Bay area. The Cryosphere, 13(2), 451-468.

Hochheim, K. P. and Barber, D. G. (2014), An update on the ice climatology of the Hudson Bay system, Arct. Antarct. Alp. Res., 46, 66–83.

Saucier, F., Senneville, S., Prinsenberg, S., Roy, F., Smith, G., Gachon, P., Caya, D., and Laprise, R.: Modelling the sea ice-ocean seasonal cycle in Hudson Bay, Foxe Basin and Hudson Strait, Canada, Climate Dynam., 23, 303–326, 2004.

Sigmond, M., Reader, M.C., Flato, G.M., Merryfield, W.J. and Tivy, A. (2016) “Skillful seasonal forecasts of Arctic sea ice retreat and advance dates in a dynamical forecast system”, Geophysical Research Letters, 43, 12,457-12,465, doi:10.1002/2016GL071396.

Zampieri, L., Goessling, H. F., & Jung, T.J. (2018). Bright prospects for Arctic sea ice prediction on subseasonal time scales. *Geophysical Research Letters*, 45, 9731– 9738. https://doi.org/10.1029/2018GL079394

Reply to Reviewer 2- tc-2021-282 Asadi et al. 2021 - Probabilistic Gridded Seasonal Sea Ice Presence Forecasting using Sequence to Sequence Learning

We sincerely thank the reviewer for the thorough review and excellent comments. Reviewer comments are shown in black, our response are shown in blue.

The authors present a fascinating application of machine learning techniques to better predict ice presence/absence within Hudson Bay, using ERA5 data as an input. The results show promise in helping plan shipping operations around the ice-free season, however, the clarity of these results is lost in lengthy wording. It is recommended that the authors read through the document for grammatical errors and places where the wording of sentences can be made more succinct. This article can become much more impactful and easier to read with more 'straight to the point' sentences.

Thank you for this comment. We agree an have revised the wording throughout the manuscript. We hope it is easier to read.

General comments:

- Ensure you are consistent using 'freeze-up' with a hyphen throughout the document, and choose either 'breakup' or 'break up' to use throughout the document

Thank you. This has been corrected.

- I am aware that it is difficult to phrase sentences when discussing the number of lead days and the two models, however, I found most sentences discussing these topics hard to read. For example, line 149:

'For example, the top row of Fig 1b shows the accuracy of forecasts launched in January using Basic model for forecast lead days of 1 to 90. E.g., the first top-left box in this figure (Fig 1(b)) corresponds to the average accuracy after 1 day forecast for all forecasts launched between January 1 and January 31, ending in January 2 to April 1 and the second box corresponds to average accuracy of forecasts launched between January 1 and January 31 ending in January 3 to April 2.'

I think it would be easier if you use articles when you are referencing lead days or models. For example: 'the Basic model' or 'a 1 day forecast'. This would make your sentences flow better while reading them, which would communicate your results more efficiently.

We agree the wording can be improved and will take this into account by doing a thorough revision.

- The results section has some statements that are more suited towards the discussion section, however I see your discussion and conclusion section are combined. I'm unsure if the section headers are pre-determined by the journal, but if they are not I would suggest making section 6 'Results and Discussion', and section 7 'Conclusion'. This would allow you to discuss your results more in depth as you present them, as I feel like some of your results could be discussed more in depth.

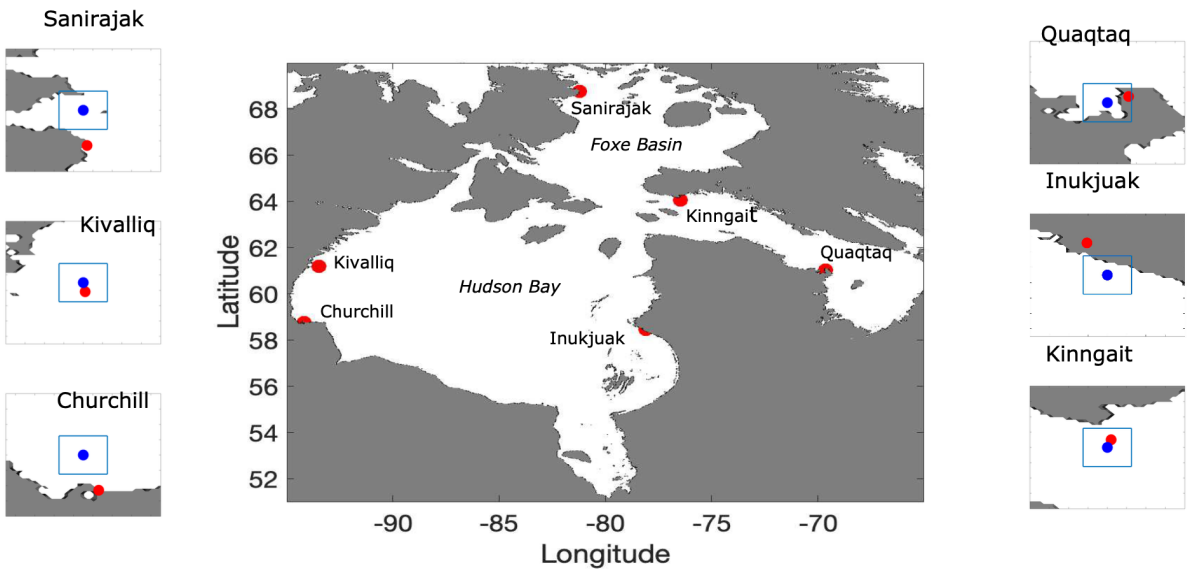
We had added a discussion section to the revised manuscript.

- Throughout the document, you abbreviate some month names and use the full name for others. You should pick one method and stick to it throughout (i.e. always abbreviate or always use the full word).

These have been fixed using the full word except for the figures, which still have the abbreviations.

- There is a comment in the specific comments regarding this, but you should include some discussion regarding the resolution of your results, and how this may impact the use of your results for port-specific operations. I am a little wary of how the land mask may impact how 'close' the pixel you use to represent the port is to the actual port in question. A figure representing this may add some clarity.

We have put together a map of the port locations, including one requested at Sanirajak (formerly known as Hall Beach). The map of the study region is shown below with the port locations shown in red. The Insets show the port location (red) and the nearest point on the model grid (blue) that is outside of the land boundary (where landmask from ERA5 is less than 0.6), in addition to a bounding box that approximates a grid cell. The model grid point near Quaqtaq is located (correctly) in the water region because the landmask from ERA5 has a low value in that region due to the low elevation.



Specific comments:

Line 3 – You may be limited on word count in your abstract, but I think it would be helpful if you stated the type of data you are feeding into your ML system to derive these predictions.

Modified

Line 3 – recommend changing to “Given the recent observations of the declining trend”

Modified.

Line 6 – recommend changing to ‘within a 7-day time period’, unless you define why a 7-day time period is ‘valid’ in the manuscript?

Modified.

Line 8 – The introductory sentence needs a little bit of work. I would recommend removing ‘northern communities’ as you do not speak of them in the rest of the introduction. Maybe focus more on the topic of shipping and why ice forecasting is vital for shipping operations in this introductory sentence. OR add in reference to communities, and why they rely on ice.

Thank you. We have changed to “Sea ice presence is an important variable for shipping operators in the Arctic and surrounding seas as it poses a significant hazard to their operations. For ships with little or no ice-breaking capability, the timing of freeze-up and break-up defines the period over which shipping operations can be carried out. For ships with some ice-breaking capability, the predicted ice cover along a proposed shipping route

provides information on transit time and is also required for accurate weather forecasts in ice-covered regions."

Line 12 – Could you expand on what you mean by 'Typical approaches are usually statistical or dynamical in nature.'? Maybe add a reference to examples of these? I see that you go more in depth in the next paragraph into dynamical forecasts, but what about statistical like you mentioned earlier?

We have added the following text to the introduction: Typical approaches are usually statistical or dynamical in nature. Statistical models have included multiple linear regression (Drobot et al., 2006), or Bayesian linear regression (Hovarth et al., 2020), whereas by dynamical approaches we are referring to those that use a forecast model solving the prognostic equations governing evolution of the ice cover (Askenov et. al., 2017, Sigmond et al. 2016). An excellent overview is given in Guemas et al. (2014).

Line 15 – I would recommend splitting this up into two sentences, breaking it up at one of the commas Modified.

Line 16 – remove 'the summer of' before 2008, as you have already indicated that this study was in the spring and summer Modified.

Line 18 – I am not too sure what you mean by 'skill'. Do you mean the forecasts ability to predict ice? There may be a better way to word this to avoid ambiguity. We agree "skill" was not specific enough.

We have changed this to "A comparison between pan-Arctic and regional forecast skill was carried out by (Bushuk et al. 2017), where skill was assessed using the anomaly correlation coefficient (ACC) between sea ice extent derived by applying a threshold to an ensemble-mean sea ice concentration and sea ice concentration from passive microwave data, and detrending both. It was shown that the ACC of seasonal forecasts in specific regions was dependent on the region and forecast month."

Line 20 – It might be nice to list some environmental controlling factors in brackets, like: (i.e. wind speed and direction, tides)

Line 24 – Recommend to change to 'Both of these approaches determine…' Modified

Line 28 – Change to 'composed of sea ice concentration data…'Modified

Line 30 – Remove 'good' Modified

Line 30 – Would help the reader if you included where the mean September sea ice extents from 2017 came from (ice charts? Passive microwave data?) Done

Changed to “Their predictions were in agreement with the mean September sea ice extent from 2017 where the sea ice extent is the total area in a given region that has at least 15\% of ice cover, according to passive microwave data.”

Line 43 – ‘calibrated probability of ice’: presence or concentration? Presence - thank you

Line 64 – Need to define ‘SST’ Modified

Line 65 – Doesn’t ERA5 have a 31km resolution? I would state this plainly so the reader knows what resolution your results are.

This information had been added to the manuscript (lines 52, 92 and 146)

Line 74 – Would recommend shuffling around this sentence: ‘Shipping traffic is also generated by mining, fishing, tourism and research activities, being mostly confined to the ice-free and shoulder season’. This part has been revised.

Line 84 – ‘In Seq2Seq learning, which has successful applications in machine translation’ Modified

Line 87 – Recommend to spell out ‘two’ Modified

Line 88 – suggest removing ‘part’ Modified

Line 92 – In line 54 you use the double wavy equal sign, but here you use a single wavy line. I would recommend picking one and keeping it consistent throughout. Modified

Line 94 – Recommend to change to: ‘The encoder section of the Basic model takes the last three days of environmental conditions as an input’ Modified

Line 97 – Remove ‘so as’ Modified

Line 99 – May be better to spell out ‘LSTM’ in full form Modified

Line 101 – Recommend rewording the last sentence for clarity: ‘The output to the encoder is a single raster with the same height and width as the input, but a higher number of channels to represent the fully encoded system state.’ Modified

Line 115 – Remove ‘so as’ (try and write sentences as simply as possible, i.e. with as little unnecessary words) Modified

Line 128 – Just verify that your quotation is facing the correct way before ‘April’ Modified

Line 137 – How did you determine what learning rater and momentum to use? We used the default Keras stochastic gradient descent (SGD) optimizer parameters (learning rate = 1e-2, momentum = 0.9). We also used a learning rate decay of 1e-4 and L2 regularization of 0.0003 with clipnorm=True. These parameters were determined as part of an initial hyperparameter search carried out at the beginning of the study when the compatibility of the data and the model architecture are investigated.

Line 142 – Suggest to remove 'coming', or replace with 'derived' Modified

Line 151 – I would recommend changing the formats of your dates here: 'forecasts launched between 1-31 January, ending in 2 January to 1 April, and the second box corresponds to average accuracy of forecasts launched between 1 – 31 January, ending in 3 January to 2 April. Changed to "forecasts launched between January 1 to 31, ending in January 2 to April 1, and the second box corresponds to average accuracy of forecasts launched between January 1 to 31 ending in January 3 to April 2."

Line 154 – This sentence needs a lot of work: suggest removing 'very' and changing 'on January' to 'of January'. As well, are you indicating that the accuracy is close to 100% for both January and the span of January – March (this is not clear)? It would be helpful if you stated the actual accuracies.

Thank you. The accuracies are close to 100% for lead days from the beginning of January to the end of March. We will revise the text to include this.

Line 155 – This sentence struggles with the same structural problems as the first, I would recommend rewording to something like: 'In contrast, for forecasts at the beginning of the open water season (June and July), the climate normal struggles to accurately capture the ice cover for lead times of 1 to 50 days likely due to inter-annual variability and the impact of climate change'. You might also want to indicate what climate change has to do with this (i.e. 'lengthening of the open water period due to climate change')

Line 165 – This sentence also needs to be reworded, I have underlined grammatical errors: 'Using additional climate variables for the input of the Augmented model is showing its impact here where in the periods that Basic model is worse than climate normal (Fig 1d), the Augmented model has better accuracy and is closer accuracy to climate normal.

We will use clearer wording in the revised manuscript.

Line 169 – Double check if it should be 'the climate normal' or 'climate normal' Checked

Line 179 – Spell out 'April' fully, as you have spelled out every other month Modified

Line 202 – ‘Observations’ should not be capitalized Modified

Figure 4 – Include units for Latitude and Longitude, and capitalize the words in your legend Modified

Figure 5 – Units for lat and long Modified

Line 212 – ‘Figure 5 and 6 show the overall…’ Modified

Line 215 – ‘The freeze-up accuracy maps at Fig 5 show that except the Basic model’s prediction at 30 lead day (Fig 6b), other maps are showing similar patterns of accuracy.’ This sentence needs reworking – would recommend flipping the sentence, so you are presenting the positive results first, then adding on the Basic model’s prediction after. Modified

Line 221 – ‘compared’ instead of ‘comparing’ Modified

Line 222 – Capitalize ‘fig 6a’ Modified

Figure 6 - Units for lat and long Modified

Line 227 – I would recommend changing all of your dates to the format: ‘1 Oct to 31 Jan’. This is a more standard way of presenting dates and is more simplistic. We have revised the dates.

Line 233 – ‘Compared’ instead of ‘comparing’ Modified

Line 234 – Change to ‘its accuracy over the breakup season…’ Modified

Line 235 – Since you discuss the break up at three sample ports, and present the results in Figures 8 and 9, I think it would be important to include a map of these three locations, indicating which pixels you use to extract this data. I am curious how the land mask affects the data, i.e. how close are the pixels you are using to the actual port? Since you are using 31km ERA5 data, I would suspect that the pixel you chose to represent each port is actually a distance away from the actual dock. In the end, I guess I am a little wary of how applicable your results are to local communities, as they are likely more impacted by ice break up on a smaller scale along the coast (for hunting and travel), whereas shipping operations are more concerned of the large scale ice break up along shipping corridors. Some discussion of how the scale of your results impacts how they are used by different groups may help address this.

We have included a map of the port locations, indicating the locations of the pixels used to extract the data. As the reviewer has pointed out these locations are a distance from the actual port. For this reason, and also because i) atmospheric conditions represented from

ERA5 would be different from those at the actual port locations ii) we do not have a complete description of sea ice conditions that can represent the complexity of port conditions, our model output is expected to be more representative of offshore conditions, which is important for route planning. We have revised the manuscript to reflect this.

Line 237 – Capitalize ‘figures’ Modified

Line 242 – Would recommend moving the figure reference to the end of the sentence, and putting it in brackets OR starting the sentence with ‘In figure 8, 30 lead day predictions for freeze-up are more…’ Modified

Line 242 – Any idea why this is? I am curious why the predictions varied at the different town ports and would think a discussion of this would add to your paper. It should be noted that the range of dates covered in the x-axis (Observed dates) varies between the locations (pixels that are near the ports). For freeze-up, the narrowest range is for Churchill (approximately 1 month), whereas Quataq and Inukjuak both have a wider range (approximately 6 weeks)..

We have included in the revised version a comparison of predicted freeze-up and break-up with that from regional ice charts provided by the Canadian Ice Service. The regional charts are weekly analyses of ice cover (concentration and stage of development of the ice) that are based on manual interpretation of synthetic aperture radar (SAR) imagery in addition to other sources, such as passive microwave and visible imagery and ship reports. A similar comparison has been done in Gignac et al. (2019). We expanded our previous port locations to include polynyas. Please see section 7.2.3 of the revised manuscript.

Line 252 – If you have space in your word count, I would recommend listing the 8 variables used in the Basic model, and the other variables added to the Augmented. This would help refresh the reader’s memory as to how these two models vary. Good idea. Thank you

Figures 8 and 9 – If possible, the font size should be increased, particularly for your axis labels. This might take some reorganizing of your figure boxes – maybe you could rotate the ‘model’ and ‘day’ labels on the far left of your figures? Modified

Askenov Y., Popova E.E., Yool, A., Nurser, A.J., Williams, T.D., Bertino, L. and Bergh, J. (2017), On the future navigability of Arctic sea routes: High-resolution projections of the Arctic Ocean and sea ice, Marine Policy, 75, 300-317.

Bushuk, M., Msadek, R. Winton, M. Vecchi, G. A., Gudgel, R., Rosati, A. and Yang, X., (2017), “Skillful regional prediction of Arctic sea ice on seasonal time scales”, Geophysical Research Letters, 44, doi: 10.1002/2017GL073155.

Dirkson, A., Merryfield, W.J., and Monahan, A.H., (2019) “Calibrated probabilistic forecasts of Arctic sea ice concentration”, Journal of Climate, 32, 1251-1271.

Dirkson, A., et al. (2021) "Development and Calibration of Seasonal Probabilistic Forecasts of Ice-free Dates and Freeze-up Dates." Weather and Forecasting 36.1: 301-324.

Drobot, S.D., Maslanik, J.A., and Fowler, C., (2006) “A long-range forecast of Arctic summer sea-ice minimum extent”, Geophysical Research Letter, 33, L10501, doi:10.1029/2006GL026216.

Gignac, C., Bernier, M., & Chokmani, K. (2019). IcePAC–a probabilistic tool to study sea ice spatio-temporal dynamics: application to the Hudson Bay area. The Cryosphere, 13(2), 451-468.

Guemas et. al., (2014), “A review on Arctic sea ice predictability and prediction on seasonal-to-decadal time scales”, Quarterly Journal of the Royal Meteorological Society, doi:10.1002/qj.2401.

Horvath, S. et al., (2020) “A Bayesian logistic regression for probabilistic forecasts of the minimum September Arctic sea ice cover”, Earth and Space Science, 7, doi:10.1029/2020EA001176.

Sigmond, M., Reader, M.C., Flato, G.M., Merryfield, W.J. and Tivy, A. (2016) “Skillful seasonal forecasts of Arctic sea ice retreat and advance dates in a dynamical forecast system”, Geophysical Research Letters, 43, 12,457-12,465, doi:10.1002/2016GL071396.

Reply to Reviewer 3- tc-2021-282 Asadi et al. 2021 - Probabilistic Gridded Seasonal Sea Ice Presence Forecasting using Sequence to Sequence Learning

We sincerely thank the reviewer for the thorough review and excellent comments. We have mainly provided responses and clarifications for the detailed questions, not addressing the typos and minor comments directly, but these would be corrected in the revised manuscript and will greatly improve the quality of the manuscript. Reviewer comments are shown in black, our responses are shown in blue.

The authors present a new approach for forecasting sea ice presence in the Hudson Bay area using machine learning techniques. The study presents models which use the Sequence-to-Sequence Learning framework to predict probabilities of sea ice presence for up to 90 days lead time. The authors suggest two somewhat different models, which are applied in hindcasting experiments, where they exhibit slightly more skill than "climate normal"-predictions especially in the breakup season. The models are also evaluated for their ability to predict freeze-up dates and breakup dates.

The study has a clear motivation, is well structured, and applies new (as to my knowledge) methods in a promising way. The text is short and precise. The general setup of the experiments is described well, however, as I'm not an expert in ML, I cannot judge the parts of the paper that go into technical details of the ML process. The results are presented clearly, but I miss a broader discussion of the results and which conclusions can be drawn from them. Especially as the motivation of the study is to develop new methods to support maritime users with new operational forecast products, a comparison to existing products would be valuable. Also a short assessment/discussion of the applicability to operational forecasting is missing in my opinion.

Please find here some general comments, followed by comments related to specific line numbers. Text in quotation marks after the given line number refers to the original text of the manuscript. Text in quotation marks in the following line is my suggestion of how to replace the original text.

%%%%%%%%%%%%%%%%%%%

General comments

%%%%%%%%%%%%%%%%%%%

A) The captions of Figures 1-3 do not only explain the figure but also contain statements about the shown results. This is a new approach to me. If the journal allows for that, I would not object, but I don't think it is common style.

We did this to help the reader interpret the figures more readily. We are not aware of limitations on figure captions for this journal.

B) For Figure 4 you compare predicted ice presence with observed ice presence. The observations are calculated from SIC from ERA5 by using a threshold of 15 %, while the ice presence from the forecasted probabilities is calculated with a threshold of 50 %. As the model is based and trained on

SIC from ERA5, why do you use a threshold of 50 % and not of 15 %? Or at least the same threshold for both?

While the model is trained on SIC from ERA5, other variables are also used, and it is an ice presence probability that is the model output. A probability over a grid cell is different from an ice concentration in that it indicates the probability of an event, which here is that the SIC is greater than 15%. A threshold of 15% is chosen for SIC, which is a common value used in the sea ice community (Stroeve 2015, Gignac et al, 2019). Note that the same thresholds were used in Andersson et al. (2021) in their study of sea ice prediction using a related convolutional neural network approach.

C) I understand that the motivation of your study is (at least partly and in the long run) to improve operational forecasting of sea ice conditions in the Hudson Bay area. In line 254-257 you explain the technical advantages of the ML approach compared to standard numerical models (=reduced computational costs). The paper would benefit from also looking into the results/skill of the ML models compared to standard models. Is your approach not only faster but also better than currently used forecasting systems? Or is it so much faster that it is useful despite of a possibly lower quality? Or is it worse/better only for some lead times? It would be interesting to see how your model compares e.g. to the S2S-forecast of ECMWF (up to 60 day forecast at 1/4 degree resolution).

https://www.ecmwf.int/en/forecasts/dataset/sub-seasonal-seasonal-prediction

https://apps.ecmwf.int/datasets/data/s2s-realtime-daily-averaged-ecmf/levtype=sfc/type=cf/

Comparing your models to the climate normal is a very good and valid first step. The comparison to a numerical forecast would however be a very interesting addition.

We agree a comparison to other data, such as the subseasonal to seasonal predictions available from ECMWF, would be a good addition to the manuscript. The sea ice information at the link provided (to the S2S ECMWF, Realtime Daily Averaged data) is available for forecast lead days up to day 46, with output twice a week, over the years 2015-2021.The central database appears to be available with data at a spatial resolution of 1.5 degrees x 1.5 degrees. We were able to download the sea ice concentration at a spatial resolution of 0.25 degrees x 0.25 degrees, which seems consistent with Zampieri et al. (2018) where it is stated “The sea ice concentration fields from the S2S database are provided on a 1.5° × 1.5° longitude-latitude grid, although the sea ice models run are at higher resolution (from 0.25° to 1°)”.

At this stage, we have carried out a comparison for the years of 2016-207 between our model forecasts and those from the S2S system. To do this the same thresholds are applied to both the predicted sea ice presence (a probability greater than 50% corresponds to ice) and the sea ice concentration from ECMWF (a sea ice concentration greater than 15% corresponds to ice). The ECMWF predictions are launched twice a week (Monday and Thursday). For the comparison, the data from our system (Basic model, Augmented model and Climate Normal) are extracted for the same launch dates as those used by ECMWF. The ECMWF data was interpolated to our 31 km grid resolution using a nearest neighbor approach. The accuracy is calculated in the same way as for the other accuracy plots in the submitted manuscript. Full details are given in the re-submission.

In our earlier response to the reviewers we found very poor performance of Basic model in November due to the opening of the Kivalliq polynya in the north western portion of Hudson Bay.. We discuss this in the resubmitted manuscript, but not the degradation is less severe due to the landmask (there is less emphasis on the central part of Hudson Bay with the 0.25 degree model output in comparison to the 1.5 degree model).

D) With the Basic and Augmented models you introduce two approaches, which you compare with each other throughout the paper. However, I don't find a conclusion/discussion about which of the two models you would suggest in the end. Is it worth the effort of the Augmented model, which needs more input data, or is the Basic model sufficient for the purpose? Or are both needed, for different purposes? Maybe you can here also explain/speculate why the Augmented model is considerably worse than the Basic model in Figure 2c. [This text could extend the summary given in lines 258-265.]

Figure 2c (original submission, which is Fig 5c in re-submission) shows the Brier score difference for the two models. A Brier score of zero indicates an optimal result. With this in mind, the negative values in this plot indicate the Augmented value is better than the Basic model for most of the ice season, consistent with the difference in binary accuracy (Fig 4f). Fig 10 shows time series of the Basic and Augmented model, and climate normal, over 1996-2017. The trend lines indicate the accuracy of freeze up and break up is slightly improved with the Augmented model, in comparison to the Basic model. Freeze-up is better predicted at 30 days with the Augmented model, which we speculate is due to the additional air temperature input (mentioned in the Discussion section of the revised manuscript).

E) For your hindcasts you use input data from ERA5, which is a reanalysis product that is not available in real time. Hence, when one wants to apply your method for forecasting *future* conditions, other input data need to be used. It would be good to elaborate on this topic in a paragraph in the Discussion or Outlook sections. Is it difficult/problematic to switch to other input data? Can the trained monthly models be applied if the forecast is started based on other data for the 3 historical days? And/Or what else is still needed before your models can be used for operational forecasting? This would be a good topic for an Outlook-section/paragraph.

Thank you for this comment. We do not envision it to be difficult to switch to other input data from the point of view of model architecture. However, the presence or absence of sea ice at different locations will be dependent on different driving factors. Other variables (such as sea ice thickness) may be critical. In addition, because this model is trained using ERA5 it will learn the dependencies and patterns in these data. We recommend that if one was to use input data from a different reanalysis that they fine-tune the existing weights to account for the different data dependencies in the input data (in particular consider that only a subset of model variables are used, dependences present in one subset may be partially considered in a different subset for a different model). Finally, while the model here is demonstrated in hindcasting mode, it can (and is intended) to be used in forecast mode. In forecast mode, given an input time series of three days, forecasts can be generated up to 90 days lead time.

%%%%%%%%%%%%%%%%%%

Specific comments

%%%%%%%%%%%%%%%%%

Title
###############
In the introduction you mention that the novelty is that your forecast is "spatiotemporal". Hence I wonder why you don't use this word in the title.
Good point – thank you

Abstract
###############
2-3 Would ML approaches be less important without global warming? Suggestion: Remove "Given ... global warming". Good point – thank you

4 "a daily spatial map" Isn't the clue of your study that you provide several "daily spatial mapS", namely 90 for a 90-day forecast?Good point – thank you

1 Introduction

##################

9, 10, 12, 14, 82 Be more clear on the terms "short-term", "longer term", "seasonal" and "medium-term" forecasting. Thank you for pointing this out. We have corrected the text and added more context. “Sea ice forecasting needs to be carried out at various spatial and temporal scales to address different requirements of stakeholders. Short-term forecasts (1-7 days) at high spatial resolution are important for day-to-day operations and weather forecasting (Carrieres et al., 2017), whereas longer term (eg. 60-90 day) forecasts are desired by shipping companies and offshore operators in the Arctic for strategic planning (Melia et al. 2016). In this study we are interested in these longer term forecasting methods, which we will refer to as seasonal forecasting.”

16 Maybe mention the lead-time used in Zhang et al. (2008) We have added text to describe this study. “This study used a coupled ice-ocean model forced by a year of atmospheric forcing data taken from a representative ensemble.”

21 "governing" Do the equations govern the physics? I'd suggest "describing" Modified

24 "This is a key advantage of..." Suggestion: "This disadvantage can be overcome by using..." Modified

27 "to perform" "for" Modified

29 "Their results are" "The results were" Modified

29, 31 In line 29 you write that the model predicts sea ice concentration but in line 31 you present results for ice extent. I'm not sure if one can assume that everyone knows the relation between SIC and sea ice extent. Thank you. We will define this difference in the revision.

32 "September sea ice minimum." minimum extent? minimum thickness? Modified to indicate sea ice extent

33 "This study" This is misleading because it could mean your own study. Better use "They" or "Hovath et al. (2020)". Modified

34 "was found the uncertainty" "was found that the uncertainty" Modified

38 "that is closer to what is proposed here" As we don't know yet, what you will propose, this information is not very useful here. Good point, we removed this statement.

49 "probability of ice at" "probability of ice presence at" Modified

2 Data

##################

56 "data from 1985-2017 is" "data from 1985-2017 are" Modified

65 "the following input variables" "the following 8 input variables". This helps explaining the number 8 in line 96. Modified

67 "V-Component" “V-component"Modified

67 replace "and" before "landmask" by comma Modified

3 Study region

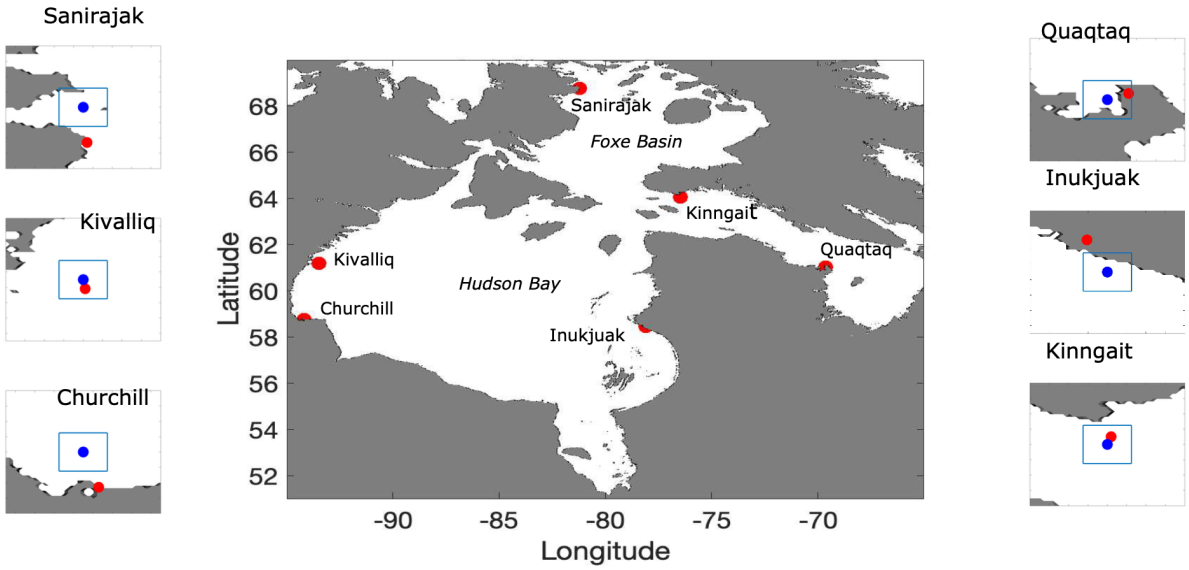
######################

Here would be a good place for a map which also indicates the ports used later on.

76 Remove the parenthesis if Foxe Basin can be shown in a map.

We agree with the above two points and will be adding a map in the revised manuscript. We have put together a map of the port locations, including one requested at Sanirajak (formerly known as Hall Beach). The map of the study region is shown below with the port locations shown in red. The Insets show the port location (red) and the nearest point on the model grid (blue) that is outside of the land boundary (where landmask from ERA5 is less than 0.6), in addition to a bounding box that approximates a grid cell. The model grid point near Quaqtaq is located (correctly) in the water region

because the landmask from ERA5 has a low value in that region due to the low elevation.



79 "Recent decades"

For me, 1985-2017 includes several "recent decades" and one could get the impression that lines 77-79 are not valid for "recent decades". So maybe consider re-phrasing "recent decades" to "In recent years" or using the term "trend"? Modified

4 Forecast model architecture

############################

85/86 "sequence of inputs"/"sequence of outputs": It would be helpful if you could mention (maybe in a new sentence) some examples for "input" and "output" for the application in this study. I guess input includes SST, t2m, winds, etc. and output is ice presence probability? Yes, that is correct, We have modified the text to reflect this."The encoder component transforms a given input (here, a set of geophysical variables such as sea ice concentration, air temperature etc.) to an encoded state of fixed shape, while the decoder takes that encoded state and generates an output sequence, here, a sea ice presence probability), with the desired length, which corresponds to the number of days of the forecast (90 days)."

86 "consist" "consists" Modified

88 Does "desired length" in your application mean number of variables or number of grid cells or number of forecasted days? It is good with a general explanation of the Seq2Seq method like you do here, but for someone not from the ML field, it would also be nice to directly get examples about how the method can be understood for the application of sea ice forecasting. In this study the desired length is the number of forecasted days, which is 90.

89-90 I first understood this sentence such that the encoder part would be called Basic model and the decoder part would be the Augmented model. Can you phrase it differently to make it more clear also for non ML-experts?

We have changed this to “using the encoder-decoder architecture described above, two spatialtemporal sequence-to-sequence prediction models are developed. These will be referred to as the “Basic Model” and “Augmented Model” and are described in Sections 4.1 and 4.2 respectively”. We are also adding a figure of the architecture of the model used, which should help clarify this (please see below)

94 "three days of environmental conditions" Shouldn't it read "environmental conditions of the last three days"? Yes thank you.

96 Maybe explain why you call the number of input variables "C":

"and C is the number of channels, in this case the total number of input variables (here 8)."

98 "sequence of extracted feature grid"

Are the feature grids what was called "environmental patterns" in the previous sentence? If so, could you use the same term? If not, could you explain how to get from one to another?

The feature grids are the “environmental patterns” referred to in the previous sentence. We will use this terminology in the revision consistently

98 "the sequence are" "the sequence is"

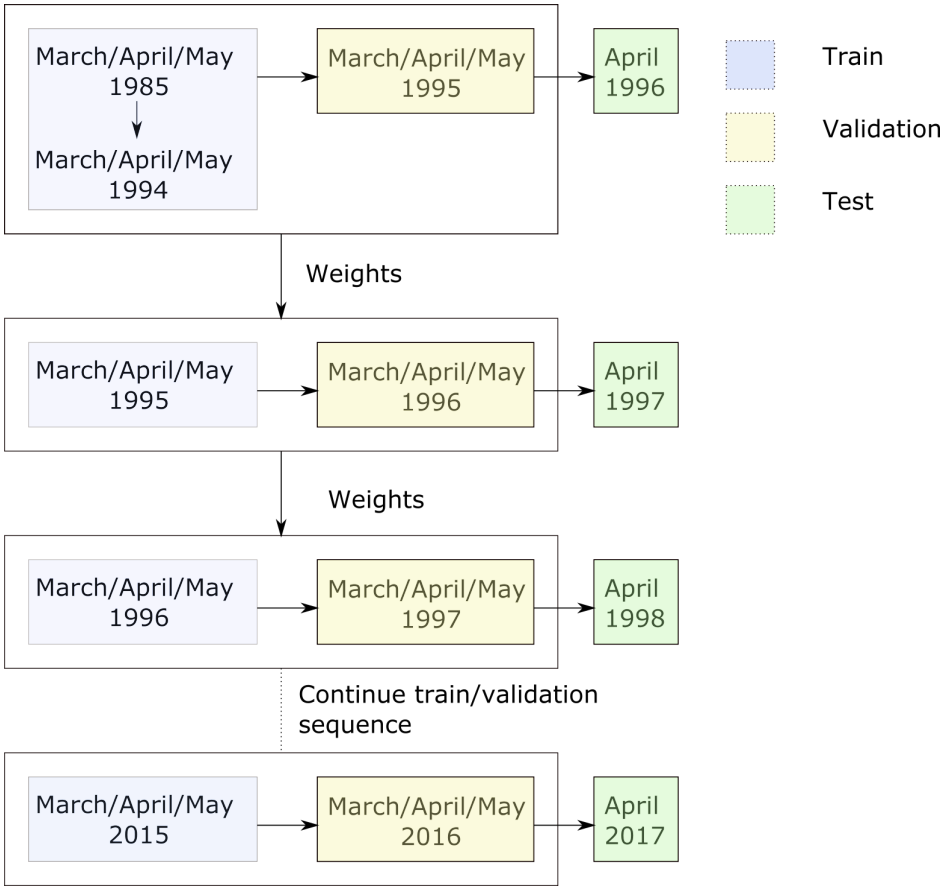
97-113 As I don't have a background in AI/ML, I unfortunately don't understand the setup of the model in detail. However, as I can follow the general concept, e.g. what is input and what is output, I think it is OK to keep the text as it is if the targeted audience is AI/ML experts more than sea ice modellers.

We will be adding more description and figures to the manuscript. The other reviewers have suggested this as well. We have two figures in mind, one describing the method in terms of data preparation and train test sequence, and another that describes the model architecture in more detail. Preliminary figures are shown below.

The first figure shows the train/test sequence. It is used to describe the following text from the manuscript,

For each month of a year a separate model is trained on data from the given month as well as the preceding and following month. For example, the ’April model’ is trained using data from March 1 to May 31. This monthly model is initially trained on data from a fixed number of years, chosen to be 10 years. After this initial experiment, to predict each following test year i, using a rolling forecast prediction, the model from year i−1 is retrained with data from year i−2 and also, data from year i−1 is used as validation for early stopping criteria and to evaluate the training performance. For example, if the initial model is trained on 10 years, data from year 11 is used as validation and first

predictions are launched at year 12. The model for year 11 is initialized with weights from the 10-year model and retrained with data from year 10, validated on year 11 and predicts year 12. The model for year 12 is then initialized with weights from the year 11 model, retrained with data from year 11 and validated on year 12 to predict year 13. This process is used to produce forecasts of sea ice presence for years 1996 to 2017.



The second figure is also shown below. The upper panel shows the overall architecture(described on lines 98-105 of the submitted manuscript, modified here).

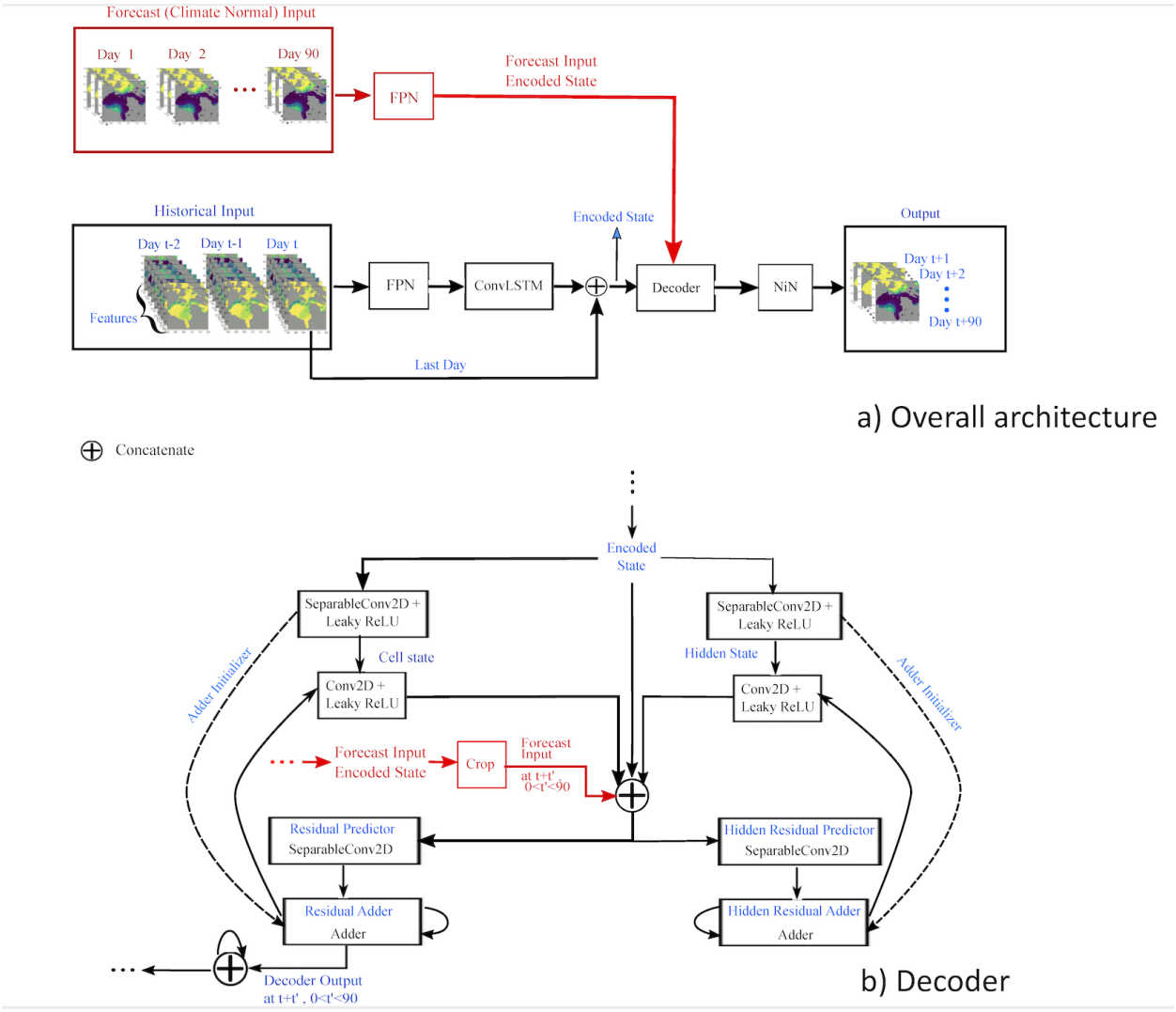
The overall architecture is shown in the figure below (panel a). The encoder starts by passing each daily sample through a feature pyramid network (Lin et al., 2017) so as to detect environmental patterns at both the local and large scales. Next, the sequence of feature grids extracted from the feature pyramid network are further processed through a convolutional LSTM layer (ConvLSTM) (Hochreiter and Schmidhuber, 1997; Xingjian et al., 2015), returning the last output state. This layer learns a single grid representation of the time series that also preserves spatial locality. Finally, the most recent day of historic input data is concatenated with the ConvLSTM output. The encoder provides as output a single raster with the same height and width as the stack of raster data input to the network, but with a higher number of channels such as to represent the fully encoded system state. The final encoded state is then fed to a custom recurrent neural network (RNN) decoder that extrapolates the state across the specified number of time-steps. It takes as input the encoded state with multiple channels and as output produces a state with the same height and width as the input

over the desired number of time-steps in the forecast (here 90 days). Finally, a time-distributed network-in-network (Lin et al., 2013) structure is employed to apply a 1D convolution on each time-step prediction to keep the spatial grid size the same but reduce the number of channels to one, representing the daily probabilities of ice presence over the forecast period (e.g. up to 90 days).

The lower panel shows the decoder (described on lines 106-113 of the submitted manuscript, modified here)

The custom RNN decoder, as is common of many RNN layers, maintains both a cell state and a hidden state (Yu et al.,2019). First, the initial cell state and hidden state are initialized with the input encoded state. Then, at each time-step and for each of the states, the network predicts the difference, or residual, from the previous state to generate the updated states using 2D depthwise separable convolutions (Howard et al., 2017). Depthwise separable convolutions are chosen to preserve the time dimension of the input, which the convolution operates over the two spatial dimensions. The output of the decoder section is the concatenation of the cell states from each time-step (unrolling of the learned time sequence).

The red portion shown in the figure below corresponds to the additional components required for the Augmented model (described on lines 115-122 of the submitted manuscript)



115 I miss a sentence about why you suggest an additional model. What is the (expected) problem with the Basic model or which benefits do you expect from the Augmented model?

The Augmented model was not developed to address a specific problem with the Basic model. It was done to enforce the climate normal, which can help the model generalize, meaning produce better forecasts over a wider range of conditions.

116 "(e.g., 60 or 90 days)" remove comma Modified

117 "t2m, u10 and v10" Modified

Why do you use exactly these variables? I could imagine that climate normals of e.g. sea ice concentration or sea surface temperature could also be beneficial to correctly predict sea ice presence.

These variables were chosen because of their availability in both historical data set, and real time (for this application, through the Meteorological Service of Canada GeoMet platform). Since this branch of the network ‘augments’ the core model, it was desired to keep this flexibility for future development as our computing infrastructure is designed to connect with GeoMet.

5 Description of Experiments

##############################

124 "required" "requires" Modified

125 "assess" "assesses"Modified

130 "the model from year i-1" "the model for year i-1"Modified

130 End the sentence after "i-2" and start a new one.Modified

136 and 140 Are "ML models" (line 136) and "neural network model" (line 140) something different? If not, use the same word. Thank you - we have changed this to be more consistent.

142 "3 months of year" "3 months of each year" Modified

141 "thresholded at 15%" This is unclear to me. Do you apply the threshold to convert to ice presence? Yes, that is why we apply the threshold.

131, 132, 142 In line 142 you talk about "test procedure". Is this the same as "validation" mentioned above? Validation refers to the use of the year following those used for training to check early stopping criteria. The test procedure is the same as the prediction procedure, referred to on line 133.

6.1 Presence of Ice Forecasts

##########################

145 "6.1 Presence of Ice Forecasts"

Shouldn't it be "Forecasts of Ice Presence"?

147 "test set" What is this? I don't think you have introduced this term before. The test set is the set of days over which the 90 day predictions are launched.

148 How do you calculate accuracy from the binary forecast map? Accuracy is the ratio of the number of correctly classified pixels (ice or water) to the total number of pixels under consideration. For example, if we consider the accuracy of the Basic model, it is calculated as: accuracy=(TP+TN)/N, where TP is number of true positives (observation and Basic model are both ice), TN is the number of true negatives (observation and Basic model are both water) and N is the total number of points considered, which is the number of non-land points multiplied by the number of days the score is calculated over.

148, 149, 150 etc.I would put a period after the abbreviated "Fig" -> "Fig." Modified

150 "in this figure (Fig 1(b))" "in Fig. 1b" Modified

150 "the first top-left" "the first (top-left) Modified

151 "after 1 day forecast" "after a 1-day forecast" Modified

152-153 Why April 1 and April 2? Wouldn't a 1-day forecast started on January 31 end on February 1, and a 2-day forecast on February 2? Yes, a 1-day forecast started on January 31 will end on February 1. This wording pertained to a 90 day forecast, which would end April 1 if launched on January 1. We will revise the wording to clarify.

154 "month on January" "month of January" Modified

155 "consistently" "constantly" Modified

156 Mention the sub-figure number you are talking about. Modified

158 "Fig 1d" "Fig. 1d and 1e" Modified

158 "significantly" Modified

Did you do a statistical test whether it is significant? Otherwise maybe remove the word. We did not do a statistical test and will change the wording.

160 "early lead times" "short lead times" Modified

163 "Climate normal" Stay consistent with capital C or not. Modified

163-164 Comparing the Augmented model with the climate normal (Fig. 1e), I don't see an improvement for March/April. I see the Augmented model is better than the Basic model, but actually it is just 'less bad' compared to the climate normal. The accuracy of the Augmented model is not higher than of the climate normal. (You explain this later, so maybe make clear that this sentence only deals with Fig. 1f and not with Fig. 1e.

Thank you for noting this. We will revise the wording

166 "(Fig 1d)" "(dark areas in Fig. 1d)" Modified

167 Remove "accuracy" at the beginning of the line. Modified

167 "90 lead day" "90 lead days" Modified

167 The "For example..."-sentence is not complete, there is no verb. Modified

169 Consider to start a new paragraph for the Brier score. Modified

169 (related to comment for line 148): What is "probabilistic accuracy" compared to "accuracy"?

The term probabilistic accuracy was used to refer to the Brier score, while accuracy is the accuracy calculated using the ratio of true positives and true negatives to the total number of samples. We will put in an equation and revise the wording.

173 Remove "Also" Modified

174 "Pt is the model prediction" Maybe add "... of ice presence probability" Modified

174 "represents" "presents"/"shows" Modified

177 "The pattern observed" "observed can easily be mixed with observations, so maybe say "The resulting pattern" Modified

178 End sentence after "both models" and start a new one for the differences. Modified

178 "(2c)" "(Fig. 2c)" Modified

179 "longer lead days" "longer lead times" Modified

184 "For early lead days" "For short lead times" Modified

185 "lead day" "lead days" (make sure to do it consistently throughout the paper, e.g. line 211 and in caption of Figure 3) Thank you

184-187 I find this sentence is too long. Also, there is inconsistency of the used terms "forecasted probability" and "forecasted probabilities".Modified

190 When first reading the sentence, it sounds like monthly averaging would make it impossible to provide information on a map. Maybe you can clarify it with "Monthly averaged and domain-integrated accuracy values ..." Modified

194-195 To simplify the explanation of the different dates, I suggest to add the dates in the caption of the subfigures 4a-4c, e.g. "(a) 5 June 2014 (after 30 days)" Modified

194 "given data" "given date" Modified

198 "and and" "and the" Modified

200-201 I don't see in the figures that the Basic model would have "increased ice presence probability in the northern part of the domain". If it is important, highlight the area in the plots.

202 "Observations" "observations" Modified

201-202 "the agreement ... to be in good agreement" Too many agreements. Modified

Figure 1 and 2:

I would suggest to use a diverging colormap when differences are displayed. This makes it easy to see where zero is located and it makes it more clear which plots display differences and which plots display absolute values.This is a good idea. Thank you.

Caption of Figure 1: "Model performance and improvements": Why not call it "Accuracy"? Modified

Caption of Figure 2: "Brier score of the Basic model (a) and the Augmented model (b) as a function of lead time. Their score difference is shown in (c). Most differences are observed in breakup and freeze-up seasons."Modified

Figure 3: - It would be nice if the aspect ratio of x and y axis was 1, so that the dashed line would be at 45 degree.Modified

- for the text in the legend I suggest "xx lead days" instead of "xx Lead Day"

Caption of Figure 3:

- Would be nice to remind the reader (especially those who only look at the figures and don't read the text) that you are talking about ice presence probabilities/frequencies. Modified

Figure 4: - In order to compare probability maps with ice presence maps it would be nice (if possible) if the plots would have the same size, i.e. smaller plots for those which have no colorbar.

- I would prefer if the height of the colorbar was the same as the height of the map-plot.

- Make sure to use "Climate normal" or "Climate Normal" consistently. Modified

Caption of Figure 4:

- The figure does not illustrate "the May models" but the forecasted conditions. Hence a suggestion for rephrasing: "Ice/water distribution in the model domain as observed and forecasted by Basic and Augmented models for a forecast started on 6 May 2014 and lasting for 30 (a), 50 (b), and 70 (c) days, respectively." Modified

6.2 Assessment of operational capability

##########

209 "15 continuous days in a row" Either "continuous" or "in a row" is enough. Modified

212 It can be a bit confusing that "accuracy" here is related to freeze-up/breakup while the same word is used in section 6.1 for ice presence. (Well done in line 215.) Modified

214 "...prediction is correct." The reader has to infer that 'correct' is translated to 1 and 'not correct' is translated to 0. Yes, this is confusing. We will clarify this.

215-216 Check the grammar of the sentence. Modified

219-212 It is surprising to me that a model can have that much more skill on a 30 day longer lead time. Could you elaborate on possible reasons and/or why the Augmented model is doing a better job? (This should probably go to the Discussion section)

We assume you are referring to Fig 5, where panels b) and c) show the accuracy of freeze up for 30 and 60 days using the Basic model, and panels d) and e) show the accuracy of freeze up for 30 and 60 days using the Augmented model.

Freeze-up dates are checked from Oct 1 to Jan 31.

30-day forecasts would have been launched Sep 1 to Dec 31. These models would have been trained on data from Aug 1 - Oct 31 (for the September model) and Nov 1 to Jan 31 (for the Dec model).

60-day forecasts would have been launched Aug 1 to Nov 30. These models would have been trained on data from July 1 - Sep 31 (for the Aug model) and Oct 1 to Dec 31 (for the Nov model).

The 60 day forecasts may be better than the 30 day forecasts because the air temperature can have more of an impact for 60 day forecasts as the open water season is considered more heavily in the training data for the 60 day model (training data extends into July). Klaus et al (2014) note a dependence of sea ice extent on fall air temperatures during freeze-up in this region. Note for the Basic model the accuracy is quite high along the west coast for 30 days, which could be due to the role of wind in this region, which would be more highly correlated with freeze-up at short time scales.

222 "breakup prediction ability" Why not call it "breakup accuracy" in analogy to line 215? Modifiedfi

222 "are presented" "is presented" Modified

222 "fig 6a" Fig. 6a Modified

227 "variability" Would "interannual variability" be more clear? Modified

227 "models' accuracy" As climate normal is not really a model, I would remove the word "models'". Modified

228 "represented" "presented" Modified

228 "For each prediction... same color." Suggestion: "The respective trends are shown by dashed lines." Modified

229,233 "freeze-up season accuracy" "season" is not necessary. Modified

229 "both lead days" "both lead times"Modified

229 "breakup plots (...)" "breakup accuracy (Fig. 7c and 7d) Modified

230 Move "accuracy" directly after "2%"? Modified

232-233 Is this because the Augmented model uses climate information as input data and hence tends to predict more similarly to the climate normal than the (more independent) Basic model does?

We see more variability in model predictions for freeze-up than breakup in general, and in particular less improvement of the Augmented model in this season. This may be because climatological trends in freeze up have changed in recent years more than those for breakup.

240 "is different." "is different (i.e. different x and y axes)." Modified

236 Refer here to the (new) overview map for locations of the ports. Modified

245 "both models both lead days" "both models and both lead times" Modified

248 "as freeze-up for breakup" "for breakup as for freeze-up" Modified

Figures 5 and 6:- Add a space between (a), (b), ... and the subfigure caption text. Modified

- Why did you choose to use a diverging colormap even though the plot does not display differences?

- Remove the grid lines which are drawn around each grid cell in order to make the plot less busy.

Caption of figure 6:- End with a period. Modified

Figure 7:- Add a space between (a), (b), ... and the subfigure caption text. Modified

- You could specify "Freeze-up accuracy" and "Breakup accuracy" also in the y-labels.

- Red and green lines are probably difficult to distinguish for color-blind persons. What about using black for the climate normal and e.g. red and cyan for the two models? Thank you for this comment

Caption of figure 7: "Dashed lines" instead of "Dotted lines" Modified

Figures 8 and 9:- The little red arrows next to the dots are hard to see and the written year numbers don't allow for getting a quick overview of the distribution of the year (e.g. whether the skill is getting better or worse over time). Did you try to plot the dots using a colormap which represents the different years (i.e. colorful dots, and the "valid area" as gray)? Then you don't need the text labels anymore. We thought about this, but then decided it would be too difficult to distinguish the precise years on the map, unless we used a lot of colors. Distinguishing the years can be useful to detect the outlier years. We did change the size of the dots and year labels to make them easier to read

- "freeze-up" instead of "Freeze-up" in x-labels and y-labels.

Caption of figure 8 and 9:- Mention that each dot represents one year. Modified

7 Discussion

###############

251 In my opinion you should mention somewhere that your model is not (yet) used for forecasting future condition but rather for hindcasting. Thank you. We will clarify that while the model can produce 90 day forecasts, it is evaluated here as in hindcasting mode to demonstrate the concept

254 "where it takes" "which takes" Modified

261 "Augmented model showing better scores comparing to" "Augmented model shows better scores compared to" Modified

262 "analysis on" "analysis of" Modified

267 "less disperse" less disperse than what? We will clarify that we mean the dates for freeze-up are less disperse than those for break up.

267 "to accepted" "to the accepted" Modified

References

#################

281 "shipping" "Shipping"

293, 305, 308 Why do you cite preprints of papers that are several years old? Modified

297 "S., R., G.," The co-author is called "Graversen R" Modified

297 "high resolution" "high-resolution"

298, 322 missing volume/issue/page number

327 "W.-c." "W.-C."?

Additional References:

Bruneau et al., (2021), The ice factory of Hudson Bay: Spatiotemporal variability of the Kivalliq Polynya, Elem Sci Anthro, 9(1), doi:10.1525/elementa.2020.00168.

Carrieres, T. et al., (2017), Sea ice analysis and forecasting, Cambridge University Press.

Melia, N., Haines, K and Hawkins, E.(2016), Sea ice decline and 21st century trans-Arctic shipping routes, Geophysical Research Letters, 43(18), p. 8720-9728.

Stroeve, J., E. Blanchard-Wrigglesworth, V. Guemas, S. Howell, F. Massonnet, and S. Tietsche (2015), Improving predictions of Arctic sea ice extent, *Eos, 96,* doi:10.1029/2015EO031431

Zampieri, L., Goessling, H.F. and Jung, T.J. (2018), Bright prospects for Arctic sea ice prediction on subseasonal time scales, Geophysical Research Letters, 45(18), p. 9731-9738.