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.

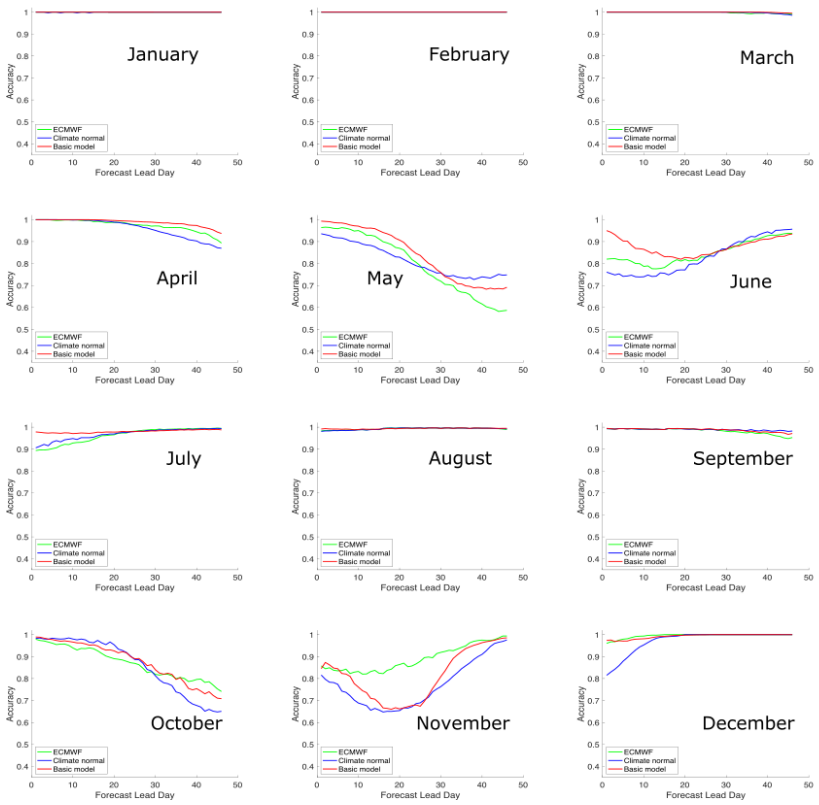
We agree a comparison to other data, such as the subseasonal to seasonal predictions available from ECMWF, would be a good addition to the manuscript. The sea ice information at the link provided (to the S2S ECMWF, Realtime Daily Averaged data) is available for forecast lead days up to day 46, with output twice a week, over the years 2015-2021. We find the spatial resolution to be 1.5 degrees x 1.5 degrees. We are not certain therefore if this is the precise data set the reviewer is referring to. We note Zampieri et al (2018) carry out an extensive comparison of various S2S sea ice forecasts for the Arctic, one of which is from the ECMWF system. In Zampieri et al. (2018) 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 year of 2017 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 (CNN-LSTM model predictions, climate normal, and observations) 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, through addition of the true positives for ice and true positives for water divided by the total number of points considered. For example, to evaluate

the Basic model true positive occurs when both the predicted probability is greater than 0.50 and the observed sea ice concentration (here from ERA5) is over 0.15, while a true negative occurs when both the predicted probability is less than 0.50 and the observed sea ice concentration is less than 0.15. The total number of points is the total number of non-land points times the number of days considered in the accuracy calculation. For each monthly model at each lead time, the number of days corresponds to the number of days in the given month (egl, 31 days for January).

Results are shown for each month of 2017 in the figure below for the Basic model. The accuracies noted for the Basic model and Climate normal are different than those given in Fig 1 of our submitted manuscript because the statistics are only calculated over non-land grid cells. Due to the spatial resolution of the ECMWF data, there are only a few points in Hudson Strait and Foxe Basin. It can be seen in the figure below that during break-up (May, June and July) the proposed method has a higher accuracy than the ECMWF forecasts, whereas during freeze-up (in particular November) the ECMWF forecasts have a much higher accuracy than those from the proposed approach. The poor performance of Basic model in November is due to the opening of the Kivalliq polynya in the north western portion of the domain. This is a large latent heat polynya that is sustained in part due to strong offshore winds. The Basic model predicts freeze-up too quickly in this region, in comparison to the observations, which here are the sea ice concentration from ERA5. The ECMWF model performs better, likely because it has a sea ice model coupled to the atmosphere. When ice starts to form this reduces the heat exchange from the ocean to the atmosphere, and the rate of ice growth slows. Our approach may not have trouble representing this phenomena because the patterns learned associated with the variables used for training could correspond to ice cover. We will need to look into this in more detail, bringing in passive microwave data (Bruneau et al. 2021), keeping in mind this ERA5 sea ice concentration is based on passive microwave data, with values less than 0.15 truncated to 0.

We are currently processing other years over which the S2S forecasts are available. Note that the observations here, that are used in the accuracy calculation, are from ERA5 (consistent with what is done in the manuscript). This is relevant because our model is trained with ERA5.



Caption: Accuracy of sea ice presence for each month of 2017 for forecasts up to 46 days. The ECMWF data are from the subseasonal-to seasonal database (s2s-realtime-daily-averaged). The accuracies shown here for the Basic model and Climate normal are different than those in Fig 1 because the points considered correspond to the landmask of ECMWF, which has a coarse grid (1.5 degrees x 1.5 degrees) and therefore some regions, such as Hudson Strait, and not well represented, and the data used correspond mostly to Hudson Bay.

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 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. The Augmented model is worse for the September forecasts (60-90 days, or sea ice states from Nov 1 to Dec 30), and the October forecasts (30-40 days, or sea ice states from Nov 1 to Dec 9). This suggests the patterns between the input features and ice presence by the Augmented model are not able to represent the sea ice evolution in November. Outside of these dates there is a small improvement when the Augmented model is used. Fig 7 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.

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, hence in 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"

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

27 "to perform" "for"

29 "Their results are" "The results were"

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"

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

67 "V-Component" “V-component"

67 replace "and" before "landmask" by comma

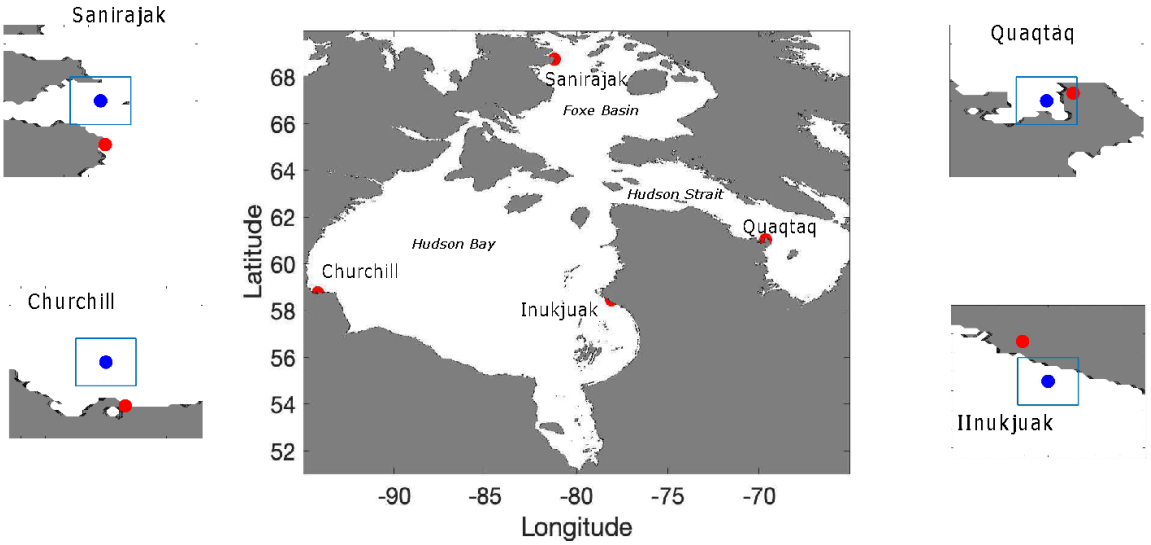
3 Study region

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

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"?

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"

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

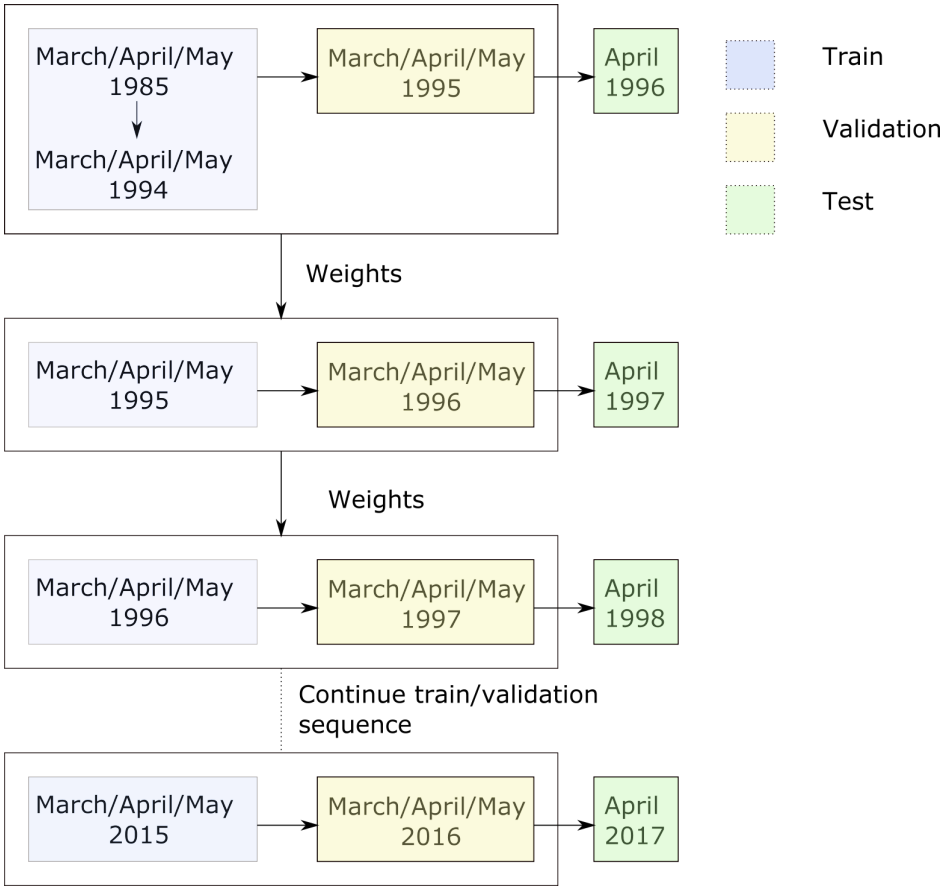
98 "the sequence are" "the sequence is"

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.



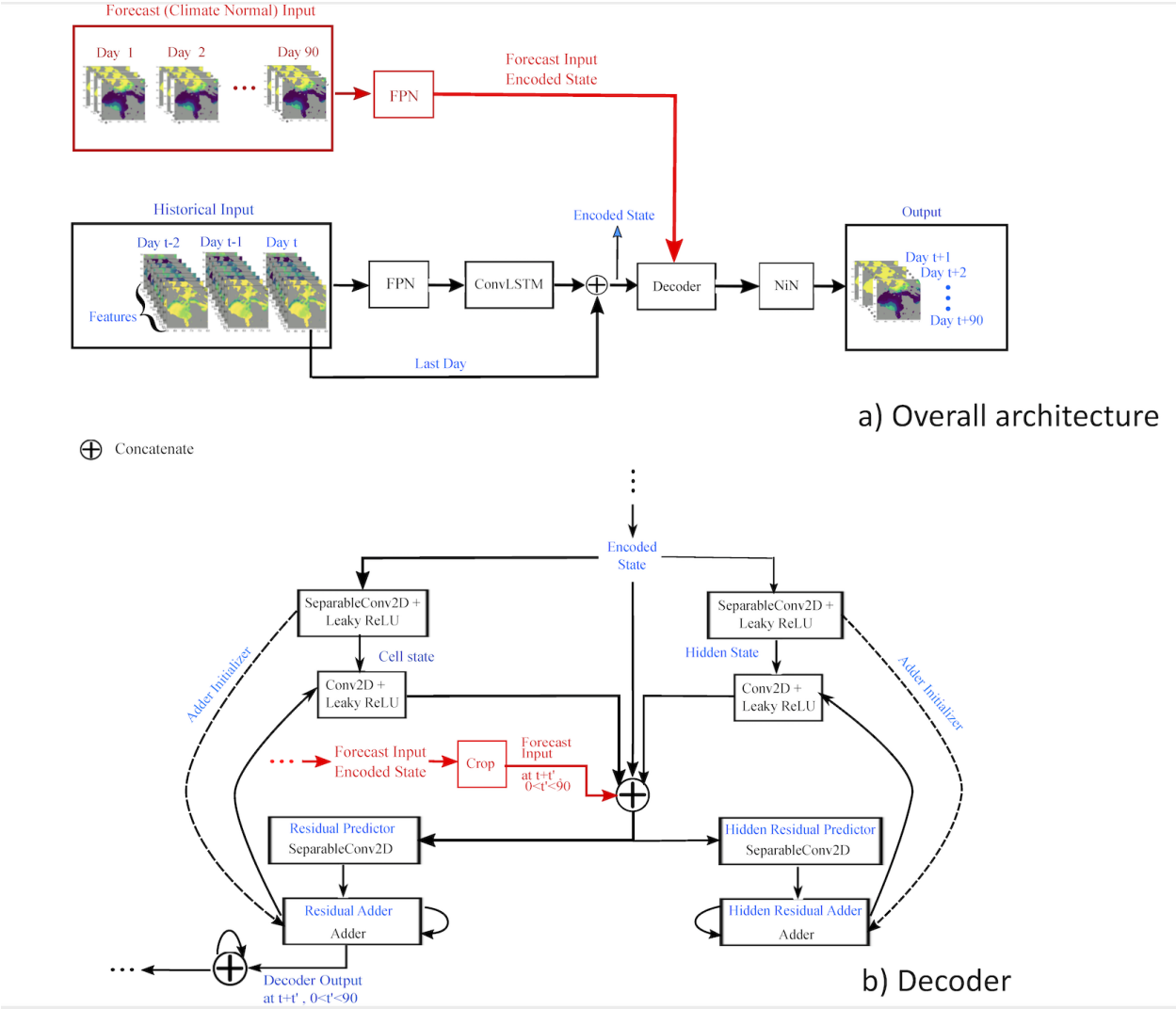
The second figure is also shown below. The upper panel shows the overall architecture(described on lines 98-105 of the submitted manuscript, modified here).

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

117 "t2m, u10 and v10"

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"

125 "assess" "assesses"

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

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

136 and 140 Are "ML models" (line 136) and "neural network model" (line 140) something different? If not, use the same word.

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

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."

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

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

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

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"

155 "consistently" "constantly"

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

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

158 "significantly"

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"

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

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)"

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

167 "90 lead day" "90 lead days"

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

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

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"

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

174 "represents" "presents"/"shows"

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

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

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

179 "longer lead days" "longer lead times"

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

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

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

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 ..."

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)"

194 "given data" "given date"

198 "and and" "and the"

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"

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

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"? =

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."

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.

- 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.

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.

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."

6.2 Assessment of operational capability

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

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

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.)

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.

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?

222 "are presented" "is presented"

222 "fig 6a" Fig. 6a

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

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

228 "represented" "presented"

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

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

229 "both lead days" "both lead times"

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

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

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. We will look into this further.

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

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

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

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

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

- 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.

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

- 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? We agree this should be modified

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

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. This is a good idea

- "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.

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"

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

262 "analysis on" "analysis of"

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"

References

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

281 "shipping" "Shipping"

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

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

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.