The paper describes the performance of the fully-coupled Regional Arctic System Model (RASM) with respect to the simulation of polynya events north of Greenland. A 42-year long simulation (1979-2020) is analysed in combination with satellite products and weather station data. Additionally, two ensembles are generated by forcing RASM with output from the Community Earth System Model (CESM) Decadal Prediction Large Ensemble (DPLE) simulations. The two ensembles, initialized in December 1985 and December 2015, are investigated with respect to precondition of winter polynya events.
Although the polynya in 1986 is included now in the revision the main part of the paper has not changed much. I am still not satisfied with the revision for several reasons and still think that the paper needs major revisions.

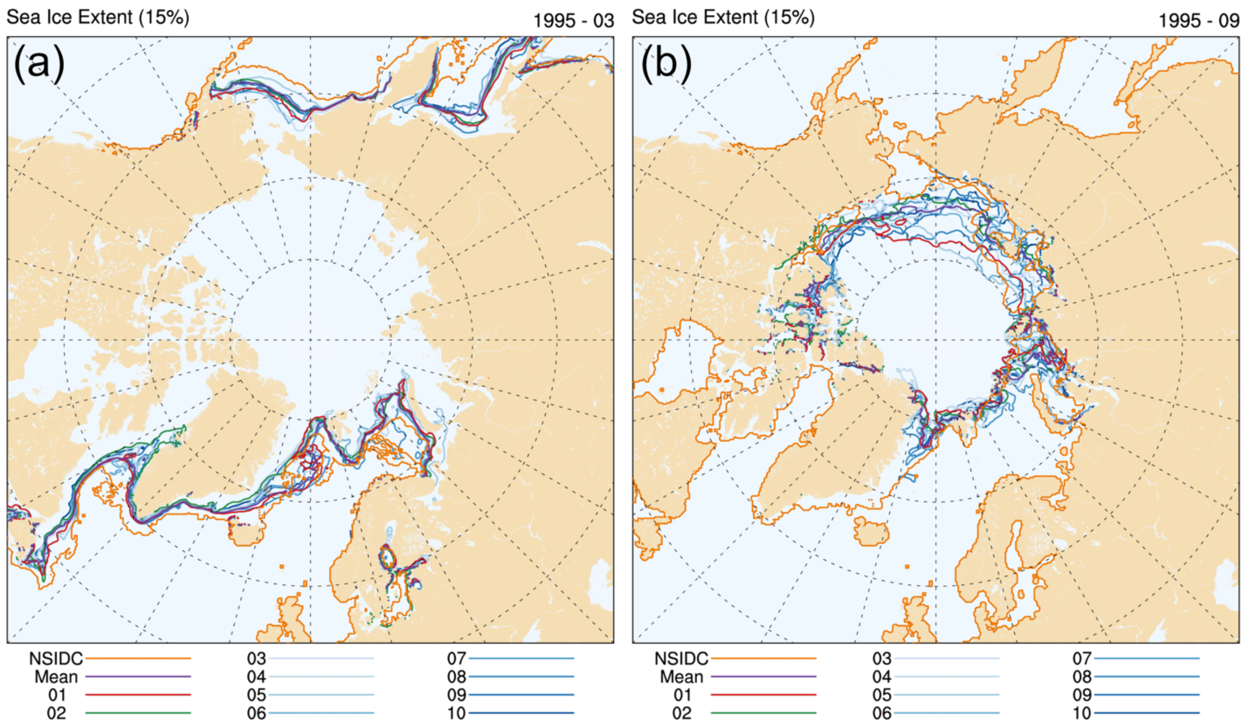
→ We thank the reviewers for taking the time to provide insightful comments that have improved our manuscript. In this revision, we thoroughly reviewed the manuscript and revised it when it is necessary based on the reviewer's comments.

[Main points of criticism]

[1]. It is still not clear to me what the scientific added value of this paper in comparison to Moore et al. (2018) and Ludwig et al. (2019) is. This should already be clearly stated in the abstract. That the polynya in 2018 is caused by mechanical redistribution is already known from Moore et al. (2018) and confirmed by Ludwig et al. (2019).
→ The abstract in the previously reviewed version of this manuscript already describes the unique contributions of this work. In this revision, we have further modified the abstract and summary to emphasize that this study examined all four winter polynya events ever observed from satellites and modeled using RASM, including 'their evolution and causality, in terms of forced versus natural variability'. To address the latter objectively, we analyzed the results from the model ensemble sensitivity studies about the influence of sea ice thickness on winter polynya development. To our knowledge, no such study has been published to date. The two papers referenced by the reviewer focused solely on the 2018 winter polynya and we have properly acknowledged their findings in the text (i.e., Lines #265-267, #280-283, and #549-552).

[2]. The two ensembles generated by RASM are not analyzed in depth. For instance, it is mainly only mentioned that "The frequency of polynya occurrence had no apparent sensitivity to the initial sea ice thickness in the study area pointing to internal variability of atmospheric forcing as a dominant cause of winter polynyas north of Greenland." This is much too general. I have my doubt, that there is any change at all in the statistics of the upper atmosphere (see below) but ANOVA is the technique to use for exactly this kind of problems (https://en.wikipedia.org/wiki/Analysis_of_variance).
→ A detailed analysis of the RASM-DPLE ensemble runs is the focus of a separate study led by one of the co-authors, Dr. Mark Seefeldt, with a manuscript in the final stages for submission. The manuscript by Seefeldt et al. (in prep.) shows that even though the mean climatic near-surface atmospheric climatic state across the 10-year simulations is similar across the 10 RASM-DPLE ensemble members, the impact on the sea ice state is indeed very different. The differences in the sea ice state are a result of the variability in the evolution and sequencing of the individual weather patterns that result in the mean near-surface climatic state. Addressing mean

climatic state statistics of the upper atmosphere only tells the story of the 10-year mean and it does not tell the story of the individual weather systems, and the sequencing of those weather systems that leads to the state of the atmosphere, sea ice, and polynyas.



Attached is a figure from the upcoming Seefeldt et al. manuscript that demonstrates the large role that variability in the evolution of the weather systems has on the sea ice state. The plot shows the sea ice extent in year 10 of the 10 1985 RASM-DPLE ensemble member simulations. Each ensemble member has a different sea ice extent, especially in Greenland and Barents seas during winter, for 1995 despite a similar near-surface mean climatic across the 10 years for the 10 ensemble members. Although not shown, ANOVA exhibits that the upper tropospheric conditions (i.e., geopotential height at 300 hPa as shown in Fig.S2a) are also similar in the 1985 RASM-DPLE ensemble. It is based on results such as are in this figure that we feel confident that the internal variability of atmospheric forcing plays a large role in the cause of winter polynyas north of Greenland.

As the reviewer suggested, we have also examined the atmospheric condition of geopotential height at the upper troposphere (i.e., 300 hPa) in the RASM-DPLE ensembles using the ANOVA test. It is found that the upper troposphere is statistically different between the two ensembles (see Table S1 and Fig. S2c). However, the polynya-causing winds are similar between the two RASM-DPLE ensembles (See also [4] regarding polynya-favorable winds). An additional figure and ANOVA table are included in the supplementary material in the revised manuscript as Table S1 and Fig. S2.

[3]. There are some typos and grammatically curious formulations in the revision that makes me wonder if the revision was done with the necessary care and if one of the native (American-) English speaking people has seen the revision.

→ All the co-authors have contributed to this work and during this revision process, we paid careful attention to typos and grammar mistakes.

[4].

(4-1) Connected to the second point. I am suspicious that the reduced number of polynyas in the second ensemble (16 versus 25) is just an artifact of the metric used (more then 10km3/day outflow of ice volume for at least three days). The mean thickness in the region of the 1985 ensemble is 3.7m and of the 2015 ensemble 2.8m. Because the metric is based on volume outflow one would expect even without any change in the wind statistics in the two ensembles a reduced number of occurrences in the 2015 ensemble, namely 25 * 2.8m/3.7m ~ 19.

→ The polynya in February 2017 was the smallest one detected from the satellite measurements (Fig. 7). Due to the lack of observational data, the threshold to define a polynya was based on the 2017 event simulated in RASM (Table 1). As the reviewer pointed out, the polynya occurrence did not account for sea ice conditions. Hence, in the revised manuscript, the columns called "the relative total sea ice volume removal (%; i.e., total ice removal ÷ (SIT*study area))" are added in Tables 1, 3, and 4, after factoring in the initial SIT. By applying an additional condition on top of the two existing thresholds, a winter polynya is stringently defined when the ice removed needs to be more than 10.8% (based on the smallest 2017 event observed; see Table 1) relative to the initial volume. We found out that 17 polynyas (reduced from 26) occurred in the first ensemble and 16 polynyas (no change) were observed in the second ensemble (Lines #425-429). The manuscript is revised accordingly but this does not affect our major findings and conclusions because a similar number of polynyas occurred under the different SIT regimes. In addition, even if we applied the 2011 polynya condition (13% ice removal), the same number of polynyas was found: 14 polynyas both in the first and second ensembles (see Lines #430-412 in the Section 4.4).

(4-2) As pointed out under (2) it has to be done a fair statistical analysis (ANOVA) to check if there is a significant difference at all between the two examples regarding the winds that cause the polynyas (there will be near surfaces differences, of course, with respect to the energy balance because the mean SIT is different).

→ The test of statistical significance is performed regarding the polynya-favorable winds between the two RASM-DPLE ensembles: 1985 and 2015 shown in Table 3 and 4, respectively. Because it is not certain how the mean wind speed and stress that cause the polynyas are distributed, a non-parametric test, i.e., the Kolmogorov-Smirnov test, is used to determine whether they are from the same continuous distribution (see Lines #417-423 in the Section 4.4). The test does not reject the null hypothesis at the 5% significance level, suggesting that the polynya-causing winds are similar between the two ensembles. In addition, the mean and standard deviation of the wind speed and stress are provided in the revised tables (see Lines #415-417 in the Section 4.4).

[5]. I mentioned already in the first revision that I find the performance of RASM in simulated the 2018 polynya disappointing. Downscaling is expected to add details to coarse resolution model results but in this case CFSv2 exists in higher resolution (about 0.2 x 0.2 degree) then WRF (about 50km). This should be clearly stated in the manuscript and that 'downscaling' is only done, if at all, with respect to the sea ice and ocean (see e.g. line 162), i.e with respect to the hindcast run the atmospheric variables are rather upscaled then downscaled. I suspect that the relatively coarse WRF resolution is responsible for the location mismatch of the observed and

simulated 2018 polynya (see Fig. 2). It would be nice to give some information about the resolution of CESM-DPLE as well.

→ As stated in the work by Saha et al. (2014) at the beginning of the second paragraph from Section 2, it is stated that the atmospheric model of CFSv2 has a spectral triangular truncation of 126 waves (T126) in the horizontal (equivalent to nearly a 100-km grid resolution) and a finite differencing in the vertical with 64 sigma-pressure hybrid layers (see Section 3.1.1). Hence, RASM-WRF (50 km) is downscaling the 100-km CFSv2 atmosphere for the 2018 winter polynya. It is often mistaken that the resolution of the model output (i.e., 0.2 x 0.2 degree) is considered the model's native resolution (i.e., 100 km). Even if the same resolution was provided by the atmospheric model in RASM, the higher resolution, and Arctic-focused, component models, and higher frequency of coupling with those component models, would provide better results than the reanalysis or global Earth system models. For example, although the RASM hindcast simulation relies on the CFSv2 atmospheric conditions, the winter SIT on 25 February 2018 is very unrealistic in CFSv2 with too thick ice in the central Arctic (Fig. 1c) (see Lines #466-470).

We have revised the model description (Section. 3.1) by adding more information on how the RASM hindcast (Section 3.1.1) and DPLE ensembles (Section 3.1.2) are designed and simulated. We also provided information about the resolution of CFSv2 (Saha et al., 2014) as well as CESM-DPLE, i.e., nominal 1° horizontal resolution and 30 vertical levels (Yeager et al., 2018) (see Lines #165-166 and #185-186, respectively). In addition, we indicated that the mismatch of the polynya location could be improved by increasing the RASM-WRF resolution in the first paragraph of Section 5 (see Lines #462-466).

[6]. There is almost no information given about the coupling of CFSR/v2 or CESM-DPLE. At least some information on the technique, the variables used and the height levels (that information is given in the discussion section, but it should be given in the methods section together with the other information).

→ Here, we assume that the reviewer is asking for more details about how CFSR and CESM data are used to drive RASM. In Section 3.1 of the revised manuscript, more information on how the RASM is set up for fully-coupled simulations (hindcast and DPLE) is given. It is clearly stated that atmosphere forcing data from global earth system models and reanalysis products are derived to provide RASM-WRF with atmospheric initial and lateral boundary conditions. Additionally, it is stated that, due to the limitation in the treatment of the model top boundary layer and stratosphere in RASM-WRF, spectral nudging of winds (i.e., u and v velocities) and temperature is applied linearly for the top half of the model domain, starting approximately above 540 hPa, with a horizontal nudging scale of up to 3400 km as described in Cassano et al. (2017), which also provides more details on information passed between WRF and the coupler in RASM (see Lines #139-199).

[7].

(7-1) Use of the AO: "The daily AO index is constructed by projecting the daily (00Z) 1000mb height anomalies poleward of 20°N onto the loading pattern of the AO" (https://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao.shtml) while the loading patterns are derived from monthly mean 1000mb height anomalies. Said that, it is clear that the daily AO is, as well as the monthly AO, a statistical mode that characterizes the Northern hemispheric state of the atmosphere but is not well suited to characterize local atmospheric states

(as e.g. in the polynya region). This is reflected by the low correlation coefficient (e.g. 0.39 and 0.45 in line 313) which means that only about 20% of the variance of the near surface temperature can be ‘explained’ by (a lagged) AO response (neither the time period used for the calculation of the correlation is given nor it is explained who is leading whom). In other words: About 80% of the variance of the near surface temperature are not correlated to the AO, i.e., independent from it which raises the question why the daily AO is considered at all. The confidence level that is given additionally is not of much help and I wonder if the auto-correlation of the time series is considered or if the daily values are considered as being independent (the auto-correlation might reduce the number of degrees of freedom in the statistical test dramatically).
→ The AO index is defined in Thompson and Wallace (1998) as a climate index of the state of the atmospheric circulation over the Arctic, affecting weather patterns more strongly in some places than in others. Since there were anomalous warming events observed prior to major polynya events in the northern Greenland region, the AO analysis is included to diagnose any potential correlations between the large-scale AO patterns and the small-scale patterns that may lead to the formation of the polynyas. However, the relationships between the observed near-surface temperature and the AO at Station 04301 in 2011 and 2017 are weak and the correlation coefficients are not statically significant when considering a lag-1 auto-correlation. Therefore, the manuscript is revised and states that “near-surface air temperature variability was not statistically correlated with the AO index” in Section 4.2.1 (Lines #311-312) and discussed in Section 5 (Lines #522-524). Also, the polynya incident related to the AO reversal is removed in the Section 6 Summary because the AO is not relevant for understanding polynya formation. If the manuscript did not include an AO analysis, there would be readers that would question why it was not included.

(7-2) In my own analysis of the polynya events 1986 and 2018 I found the plots attached below much more helpful as any correlation coefficients. They show nicely that southerly winds are caused by high pressure systems over the Barents Sea in both events, but that 2018 is connected to an SSW and 1986 not (if the SSW is causing the winds in 2018 or if that is just a coincidence cannot be answered in my opinion but obviously similar strong winds can occur without an SSW event (1986)).
→ In agreement with the reviewer’s analysis, the results in the previously reviewed version of this manuscript already emphasized that all four winter polynyas north of Greenland were driven by strong southerly winds in the Sections 4.1.3 (2018 polynya), 4.2.1 (2011 and 2017 polynyas), and 4.3 (1986 polynya) (see also Figs. 4, 6, and 9). We already discussed that SSW was not prerequisite for polynya development because the winter polynyas in 1986, 2011, and 2017 were not associated with SSW (Lines #535-538 in this revised manuscript).

[Minor point]:

[M-1] Section 3.2: Ad hoc it is not clear to me why SOM should be used instead of conventional Empirical Orthogonal Function analysis. A short explanation might be helpful.

→ It is stated that, since atmospheric wind fields might be nonlinear in nature, a linear method, such as the empirical orthogonal function (EOF) analysis, may have drawbacks in extracting nonlinear information (Hsieh, 2004) (see Lines #201-202).

[M-2] Line 235 – 240: Likely the location misfit of the polynya (Fig.3) can be attributed as well to the relatively low resolution of WRF.

→ We have addressed it and discussed that the RASM-WRF resolution needs to be improved for better representation of surface winds (see Lines 462-466).

[M-3] Section 4.1.2: I wonder why 1986 is not mentioned in this section. Are no station data available? If yes, that should be mentioned. If data are available they should be discussed as well for 1986.

→ The polynya that occurred in December 1986 is analyzed and presented in Section 4.3. Also, it is stated that no data are available for the 1986/1987 winter in Section 2.1.

[M-4] Line 250: I find the sentence "Figure 5 shows the RASM thermal sea ice surface, lateral and bottom melting terms were all negligible (< 1 cm) over the study region when integrated for the whole month of February 2018." hard to comprehend. May be better to write " Figure 5 shows that the thermal ice melting terms of RASM (at the surface, lateral, and at the bottom) are all negligible (< 1 cm) over the study region when integrated over February 2018."

→ It is revised as suggested.

[M-5] Line 270-271: "… due to the rapid ice growth during the polynya opening …". The ice is mainly growing after the opening. I suggest: "… due to the rapid ice growth following the polynya opening …"

→ It is revised as suggested.

[M-6] Line 275-276: "… and have found that the polynya development was associated with strong and persistent winds from the south-southeast." This is no new finding. I suggest: "… and have found that the polynya development was associated with strong and persistent winds from the south-southeast in agreement with Moore et al. (2018) and Ludwig et al. (2019)."

→ It is revised as suggested.

[M-7] Line 314: See main point of criticism (7). If you want to stick to the comparison with the AO: "… lagged by approximately two weeks.": I suggest to make clear that the near surface temperature is leading the AO (which means that the AO can not be causing the temperature anomalies), ie. I suggest "… while the AO is leading by approximately two weeks.".

→ It is revised as pointed out in the major criticism (please see the response to [7]).

[M-8] Section 4.3: Please check for grammatical correctness. Especially the last sentences "However, the mean turbulent heat flux was much less in December 1986 than in February 2011 even though it was a larger event in terms of polynya size and wind intensity. This is possibly due to the fact that sea ice was thicker in 1986; for example, the mean SIT was 4.4 m for 5 days before the polynya (Table 1). Due to large open water areas in December 1986, the integrated turbulent heat loss was much larger compared to the polynyas in February 2011 and 2017." are not understandable for me.

→ This section is revised by correcting errors and removing the sentences that are no longer necessary including the relationship with AO because of no observed data (thus, Fig. 9 is revised as well) and the integrated turbulent heat loss.

[M-9] Line 441-455: See major point (4).

→ This paragraph is expanded by adding the null hypothesis test using the Kolmogorov-Smirnov (K-S) two-sample test on the polynya favorable winds. Also, we introduced an additional criterion to define the polynya in RASM-DPLE by taking into account the initial sea ice thickness (please see the response to [4])

[M-10] Line 456: 'longest' → 'largest' ??? or 'longest lasting' ???

→ It is changed to "longest-lasting"

[M-11] Line 493-497: "Overall, the more frequent winter polynyas, produced in a thicker sea ice regime between the two 30-year apart ensembles, implies that changes in SIT are not significant contributors (at least up to now) to the generation of such events for this region during wintertime. Therefore, the findings support that polynyas becomes prevalent when southerly winds are more persistent and stronger in northern Greenland." See my concerns outlined in major point (4). The last sentence I do not understand.

→ The sentences are revised to clarify that the thinning of sea ice is not a significant contributor to the generation of winter polynyas. Even if the initial SIT is considered to define polynyas in RASM-DPLE, the number of polynyas is similar between the two ensembles (please see the response to [4]).

[M-12] Line 500: "… which means that the model is prescribed with reanalysis or gridded products on every grid cell.". That sub sentence is simply untrue. Correct is that forced sea ice-ocean models are one-way coupled. But the reanalysis is not prescribed as stated by the authors. 10m wind is acting via drag formulations on the ice, 2m temperature and humidity act via sensible and latent heat fluxes calculated at the surface by the model and normally downward long- and short-wave fluxes act at the surface but the net fluxes are calculated as well by the model. This sentence should be revised.

→ It is revised as suggested; forced sea ice-ocean models are one-way coupled. Although net heat fluxes are calculated at the surface in the forced ice-ocean model, they are not exchanged across the air-sea interface and do not alter the specified sea ice state.

[M-13] Line 561: The sentence "By taking advantage of an ensemble approach, the internal variability is better assessed with respect to the occurrence of such coastal polynyas during extereme events." needs justification. From which analysis presented I can deduce that with the ensemble approach the internal variability is better assessed? What is meant by 'better'? Better compared to forced sea ice-ocean models? That would be a trivial sentence, of course.

→ This sentence was replaced with the following: 'The combination of decadal dynamical downscaling with an ensemble approach allows us to increase the sample size to 100 winters for each of the two ensembles and thus quantitatively evaluate the impact of decreasing ice thickness on the occurrence of polynyas in the region.'

[M-14] Line 637: I suggest to change the sentence to: "However, the mean turbulent heat loss in the study region during the polynya in 2018 was about 61 W/m 2 (with a maximum of 124 W/m 2 at day XX), which is in good agreement with the results of Ludwig et al. (2019) based on the forced sea ice-ocean model NAOSIM (mean/maximum 40 and 124 W/m 2 , respectively).

→ It is revised as suggested.

[M-15] Line 656: 'that that' -.> 'that'

→ It is corrected.