Our comments in red
Reviewer comments in black

COMMENT TO BOTH REVIEWERS

We once again thank both reviewers for their extensive, thoughtful and constructive feedback.

The revisions requested were generally very minor, and have all been implemented. The only point that required more attention and textual revision is related to the first reviewer's question on field significance in Figure 16a (the Z500 ACC changes). The Methods section now describes two such tests that we apply, and the discussion around Figure 16 includes the results of these tests and our interpretation. These details are also included in our response to the reviewer below.

We hope that these changes are found acceptable to the reviewers, and we thank them once again for their time and efforts.

Best wishes,
Kristian Strommen (on behalf of the authors).

REVIEWER #1

Overall Thoughts

This is my second review of the manuscript. The authors did a great job acknowledging and addressing the various comments from the two reviews of the original manuscript. In particular, the authors have streamlined their discussion of the results, discussed further the summer results, and clarified some of their methodological choices. Altogether, the manuscript is stronger and clearer and remains scientifically interesting. I still think that the manuscript has the potential to be significant in the seasonal forecasting community and in research involving Arctic and mid-latitude weather connections across timescales. I still think there are a couple of areas that need some slight improvement, but this will not ultimately curtail the publication of the manuscript. Hence, I recommend that the paper be accepted for publication pending minor revisions.

Minor Revisions

1. ACC Skill Score Discussion. I am not going to belabor or get into a "back-and-forth" with the authors about the use of ACC as forecast skill. I am well aware that the community has adopted this metric to represent a form of forecast skill. If not already read and/or of interest, I encourage the authors to read Murphy and Epstein (1989) to understand the context of ACC as a skill metric.
With that, I appreciate the additional context provided by the authors on comparing

ACC "high skill" to the relatively large RMSE errors (Lines 475-477). This type of caveat reflects the spirit of my original comment – a high ACC score does not necessarily mean high forecast skill for a particular variable when compared to other skill metrics. Admittedly, for forecasting purposes, magnitudes of features could be considered more important than just pattern correlation, particularly when forecasting extremes on subseasonal-to-seasonal timescales. So, thank you for adding this text!

There are a couple of other areas in the manuscript where this concept could be emphasized as well.
a. Abstract: In lieu of the caveat above, the statement that there is a "significant increase in seasonal forecast skill" of Z500 isn't quite accurate. The improvement is shown only for ACC, not for all skill metrics; e.g., RMSE remains essentially unchanged on average.
b. Line 621: The improvement in "forecast skill" is only shown for ACC, so this would be an improvement of the ACC but necessarily of the total Z500 "forecast skill."

We thank the reviewer for the pointer to Murphy and Epstein. We agree that this makes very valuable points that have perhaps not been fully assimilated by the community. However, we are keen to not be sidetracked by this point in the present context, so we are grateful to the reviewer for accepting our usage of the word "skill". We have in any case followed your suggestions and revised the abstract to say "increase in anomaly correlation coefficients" instead of skill, and similarly for line 621.

2. Field Significance. The authors acknowledge the gridpoint-based hypothesis testing when looking for significant differences in ACC scores. However, they have not mentioned performing field significance/false discovery rate tests on the Z500 field in response to this concern. Have the authors investigated these tests? If so, what is the resulting map? The conclusions may or may not change, but this added analysis would be important for completeness.

Thanks for this comment. We had not conducted field significance tests, due to the complexity of doing this in the presence of strong spatial correlations. However, we have now tried to test for this in two different ways. We will first report the result of these two tests and then afterwards discuss our interpretation of these results.

Firstly, we applied a basic False Discovery Rate (FDR) correction based on a false-positive threshold of q=0.05 using the Benjamin-Hochberg method (https://link.springer.com/rwe/10.1007/978-1-4419-9863-7_1215). This effectively applies a blanket-correction to the entire spatial field of p-values. This had the effect of increasing the smallest p-values from below 0.05 to around 0.3. Thus, after the correction, statistical significance is not achieved at any gridpoint.

This FDR correction does not account for the spatial coherence of the patterns seen in Figure 16a. Thus, secondly, we conducted a field significance test based on Monte Carlo sampling as follows. We recomputed the difference in ACC between SPP and CTRL after randomly shuffling the time points in the ERA5 reference data, and for each such random shuffle also computed the p-values based on the Siegert et al. test. The shuffling in time ensures that the spatial correlations inherent to the Z500 field in the respective data sets is respected when computing the p-values for these shuffled fields. We carry out 50 such random shuffles, and compute statistics on the p-value fields obtained. They can be interpreted as statistics for the null hypothesis that there is no difference between CTRL and SPP (since both will have purely random correlations with ERA5 due to the time shuffling). Two statistics are computed: (1) the fraction of gridpoints passing significance, and (2) the largest area of coherent significant gridpoints found. For this latter statistic, coherence is measured empirically by joining up neighbouring significant gridpoints using scipy.ndimage.label. We then compute the values of these statistics that actually occurred in the raw (unshuffled) data sets and compare against the distributions obtained from the shuffled data. The conclusion is the same in both cases: the probability of getting the true value by chance is between 40 and 50%. In other words, even when taking into account spatial coherence we cannot establish field significance.

[Note that we only did 50 such shuffles because each case requires us to compute gridpoint correlations 3 times on a 1 degree grid, amounting to looping over ~90,000 computations each case. This was quite slow as we implemented it. However, since we are far from significance even with 50 samples it is clear the conclusion won't change by boosting the sample size further]

We interpret this as justifying the high level of caution we adopted when reporting the change in ACC with SPP. We do not, however, view it as a justification for removing all discussion on the change, for three reasons. Firstly, a large body of literature suggests that Arctic sea ice exerts a robust but weak influence on the North Atlantic jet, and SPP evidently is causing notable changes to the sea ice. Secondly, the analysis shown in Figure 16c offers modest evidence for a mechanistic pathway. Thirdly, and most importantly for the authors, a similar change in ACC was found in two other partially independent IFS experiments (a higher atmospheric resolution hindcast using Perlin noise instead of Shapiro; and the horizontal atmospheric resolution; adding stochastic perturbations to the ocean component), and even weakly in sub-seasonal forecasts with SPP. This consistency suggests that the change in ACC is likely robust in cycle 49 of the IFS, though the precise magnitude and spatial pattern of the change is hard to estimate.

In the revised manuscript, we have therefore opted to discuss the results of these field discovery tests in Section 6.2, and further emphasised the need for caution. The tests themselves are described in the Methods section.

Grammatical / Typographical Error Corrections

1. Lines 5 and 99. Please capitalize "Northern Hemisphere."

Done.

2. Line 89. "The ensemble of 50 atmospheric initial conditions are based on ERA5…" à "The ensemble of 50 atmospheric initial conditions is based on ERA5…"

3. Line 91. "an ensemble of 11 conditions are used…" à"an ensemble of 11 conditions is used…"

4. Line 96. "…the generation of the ocean and sea ice initial conditions do not make use of…" à "…the generation of the ocean and sea ice initial conditions does not make use of…"

We made the three corrections above.

5. Line 165. The parameter name "rn_csrd" should be "rn_csrdg" based on Table 1 (or conversely, Table 1's labeling is incorrect).

Thanks for catching this; the text was incorrect, and we have adjusted it.

6. Line 200. "The increased SIC in the Chuckhi and Bering seas in DJF are not significant…" à "The increased SIC in the Chukchi and Bering Seas in DJF is not Significant…"

7. Line 328. Double period after "absent."

8. Line 335. "The linear correlation ranges between −0.5 to −0.7 in the two seasons…" à"The linear correlation ranges between −0.5 and −0.7 in the two seasons…"

9. Line 421. Remove "is" after "SIT."

10. Line 423. "An asymmetric impact associated with irreversible processes such as ridging are also…" à "An asymmetric impact associated with irreversible processes such as ridging is also…"

11. Line 452. "Greeenland sea" à"Greenland Sea"

12. Line 471. Eliminate "the" between "surprising" and "there."

13. Line 503. "strenghtening" à"strengthening"

14. Line 560. "heatfluxes" à"heat fluxes"

15. Line 602. Missing the closing parenthesis after "S9."

We fixed all these.

Reference

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model *verification. Monthly Weather Review, 117, 572–582, https://doi.org/10.1175/1520-0493(1989)117%3C0572:SSACCI%3E2.0.CO;2*

REVIEWER #2

This is the 2nd review for this paper, which is an overview of using stochastic parameter perturbations (SPP) in the SI3 sea ice component of ECMWF's seasonal forecast system, SEAS6 (latest version). The manuscript describes the SPP scheme, based on the Storto and Andriopoulos [2021] STOPACK scheme, and then goes on to describe the schemes impact on spread within the sea ice and in the atmosphere. Finally, it describes the impact on various skill diagnostics for both sea ice and the atmosphere, further detailing some of the possible teleconnection impacts in the atmosphere.

I particularly commend the authors on their response to the reviewers (both) and improvements to the paper with regard to parts of the manuscript describing impacts to the atmospheric forecast (increased skill in auto-correlation of Z500 field). Although somewhat secondary to the main goal of the paper, which was to describe the impact of the stochastics perturbations to the spread, mean state and error of the sea ice state, the authors have improved (an already good) description of the atmospheric effects (sections 4.2/6.2) due to the stochastic effects, guiding the reader through the effects and the authors' interpretation/hypothesis of the dynamics involved. As commented by the other reviewer, this is likely to be "one of the top results of the paper." They have done this in a concise fashion that then allows for their comprehensive description of the main purpose of the paper, the aforementioned description on the sea ice state. I was very interested in the authors' suggestion that the sea ice perturbations are especially impactful to regions of possible Rossby wave generation, which should open up areas for the authors and others to further investigate in the future.

I would now strongly recommend the paper for publication, subject to some minor technical revisions which I will list simply in order of appearance.

We thank the reviewer for their positive comments and appreciation. We thank them once again for their very helpful first review, and for taking the time to give the revisions a thorough reading as well.

Comments (in order of appearance)

1. (l. 32) I would suggest to add the stochastic ocean perturbation work of Brankart et al. [2015], as that is the stochastic code included as the default in NEMO4.

Thanks, we agree this was an oversight and have added the reference.

2. (l. 67) Optional: "Impacts on Antarctica are included in the Supporting Information." I suggest this only because of the reply to the other reviewers comment: "Impacts on Antartica are included at the end of the Supporting Information."

We made this change.

3. (l. 101) "there can be substantial uncertainty in the observations (of sea ice concentration), and different products" This is somewhat of my own doing, as it is in response to my comment from the original review. I would be careful of the wording here, the observations (of satellite radiance) are to my knowledge not particularly uncertain. What is uncertain, is the sea ice concentration analysis obtained from the satellite radiances and other remotely sensed observations of sea ice concentration, as there is uncertaintly stemming from the size of the (SMMR/SMMI/SMMIS) satellite footprint, the algorithm used to convert radiance into sea ice concentration, and finally exactly what the satellite is sensing – 1 fraction of sea ice to open ocean [ice concentration], or fraction of exposed sea ice/snow to melt pond cover [1 minus melt pond fraction] – and likely other sources of uncertainty. I would suggest "there can be substantial uncertainty in the production of sea ice concentration analysis obtained from the remotely sensed proxies (e.g. satellite radiances) of sea ice concentration, different products" (it works with/without the "and")

We have revised the text to use a slightly shortened version of your suggestion: "For observations, there can be substantial uncertainty in the production of sea ice concentration analysis obtained from remotely sensed proxies (e.g., satellite radiances), and different products may therefore have non-trivial differences"

4. (l. 120) Comment – no change: "Informally, the CRPS is the generalisation of the mean-squared error." Technically, CRPS is the generisation of absolute error (2nd term of eqn. 3 disappears if N=1), but they did prefix the comment with "informally," so I am happy with it as it is. Additional aside: From the eqn. 3 formulation of CRPS it is not inherently obvious that the value is positive definite and bounded below by zero (requiring xt,k = yt for all k, i.e. a perfectly deterministic prediction), perhaps it would be advisable to say "Lower values of the positive definite score indicate better performance.

We revised to say "It is positive definite, and lower values indicate better performance". We agree this is not the formulation of CRPS which is most helpful for clarifying its properties, but it's a form which is very succinct to write down, which is why we used it.

5. (l. 126) "and then subtracting this (the bias) from every ensemble member before computing the CRPS." In response to my major comment # 3 from the original review, this is now done for CRPS and RMSE (and inherently done for correlation).

Whoops, thanks for catching that. We now say it happens for RMSE as well.

6. (ll. 156–160). Comment only: Thank you for providing some additional information concerning the Perlin Noise filtering of spatial lengthscales. I likely did not follow much of the explanation after "interpolation," but the additional information should allow the reader (and perhaps even me if given enough time to digest it) to better understand how the spatical *filtering is achieved and at what (approximate) lengthscale this is done for. And I did learn a new word — lacunarity; even if I would struggle to actually use it.*

Yes, sorry about this. In an ideal world we would have had equivalent experiments using Shapiro noise, side-stepping the need to suddenly introduce this additional information with such extreme brevity. We appreciate your patience with the less-than-ideal world we put before you.

7. (ll. 161–162) I believe you are missing a “to”: thus comparable but not identical to the Shapiro filter length scale.

Thanks, we fixed this.

8. (ll. 204–205) Comment only: You probably should have gently informed me in your response to the review that I did indeed assume the incorrect measure of thickness – ice present average thickness, not grid-cell average thickness.

Apologies for any inconvenience caused!

9. Figure 9. I am afraid you may have misinterpreted my comment (minor comment #16) about the masking of values. The temporal variability of sea ice in Hudson (and Ungava) Bay (and Labrador Sea and North Pacific – but that is not as obvious) should be present thoughout the bay – and not artificially truncated/masked south of 60◦N. I had assumed it was masked below 60◦N (regardless of whether sea ice is present there at some point in the year) so as to correspond to the integrals computed for figure 9a.

You’re right, we did misunderstand. We have revised the plot so that the domain extends all the way down to 40N, both for the spatial maps and when computing the Arctic average (with gridpoints with no ice masked as before). Note the line plots (Fig 9a) have only fractionally changed due to the region 40-60N contributing a small amount to the total Arctic average. But the spatial maps at least are no longer artificially truncated.

10. (l. 328) There are two (..) periods at the end of the sentence.

Fixed.

11. (ll. 421–422) “Regarding the increase in the mean SIT is, this is probably related to its own distributional asymmetry:” The sentence structure seems bungled — perhaps a stray “is” after SIT?

We rephrased the whole sentence to the following: “We have so far only discussed SIC, but we believe the increase in the mean SIT is also related to a distributional asymmetry: SIT is bounded below but not above, and hence random perturbations will tend to push the SIT towards higher values.” We hope this is less bungled.

12. (ll. 514) “Changes to SIC in this region could plausibly therefore also drive variations” It is your choice, but I feel the “therefore” is somewhat extraneous.

We removed “therefore”.

13. (ll. 548–549) Comment only (and perhaps really only a pet peeve): “ensemble spread may already be saturated in IFS, and that further improvements to spread-error ratios require a reduction of the error rather than an increase of the spread.” Or perhaps unphysically too large atmospheric perturbations could be reduced as more physically representative perturbations to the sea ice and ocean subsurface are added?

Thanks for raising this point! We agree this is a potential blind spot, and it would be very interesting to understand this better.

14. (ll. 749–755) Likely due to software (BibTeX) beyond the authors’ control, but I found the ordering of the 1st author’s 1st authored references confusing (encountered when searching for Strommen et al. [2022]). The chronicological order is 2024, 2019 and 2022 — but even if the ordering is first by alphabetic order of 2nd author (and subsequent authors) before date, should not Christensen come before Cooper. Perhaps a matter for the typesetters.

We agree that the ordering is counterintuitive, but as you guessed, it is what BibTex produced automatically, which makes it awkward to change. We will leave it to the typesetters to deal with.

References

*J.-M. Brankart, G. Candille, F. Garnier, C. Calone, A. Melet, P.-A. Bouttier, P. Brasseur, and J. Verron.
A generic approach to explicit simulation of uncertainty in the NEMO ocean model. Geoscientific Model
Development, 8(5):1285–1297, 2015. doi: 10.5194/gmd-8-1285-2015. URL https://gmd.copernicus.
org/articles/8/1285/2015/.*

*Andrea Storto and Panagiotis Andriopoulos. A new stochastic ocean physics package and its application
to hybrid-covariance data assimilation. Quarterly Journal of the Royal Meteorological Society, 147(736):
1691–1725, 2021. doi: https://doi.org/10.1002/qj.3990. URL https://rmets.onlinelibrary.wiley.
com/doi/abs/10.1002/qj.3990.*

*K. Strommen, S. Juricke, and F. Cooper. Improved teleconnection between Arctic sea ice and the North
Atlantic Oscillation through stochastic process representation. Weather and Climate Dynamics, 3(3):951–
975, 2022. doi: 10.5194/wcd-3-951-2022. URL https://wcd.copernicus.org/articles/3/951/2022/.*