Our comments in red
Reviewer comments in black

COMMENT TO BOTH REVIEWERS

We thank both reviewers for their detailed and thoughtful feedback on the manuscript. Our revision attempts to address all the raised points. Below we go through all the specific points one by one, but we first wanted to make a general comment about the choices we made in revisions.

Reviewer #1 expressed a feeling that the manuscript is long and dense, a sentiment we broadly agree with. Reviewer #2 on the other hand made a number of comments asking for additional information or analysis. This has inevitably resulted in some tension when deciding how to revise the manuscript. Our compromise has been to address Reviewer #2's questions as much as possible, but as briefly as possible. We have focused on clarifying the already existing analysis and text, and avoided as much as possible introducing additional analysis or figures. This is both to avoid the paper becoming excessively long, but also because the follow-up paper focused on sub-seasonal forecasts will fill in some of the gaps as far as additional analysis is concerned. We have made this last point more clear in the revised manuscript.

We hope both reviewers will find this compromise acceptable.

Best wishes,
Kristian Strommen, on behalf of the authors

REVIEWER #1

Summary

This manuscript presents an implementation and evaluation of a stochastically perturbed parameter (SPP) scheme applied to the sea ice component of ECMWF's Integrated Forecast System (IFS). The authors assess the impact of this scheme on ensemble spread, mean state, and seasonal forecast skill for both sea ice and atmospheric variables across the Northern Hemisphere (NH) during winter and summer. The main finding is that applying the SPP scheme enhances ensemble spread for both sea ice concentration (SIC) and sea ice thickness (SIT) by approximately 10% relative to unperturbed forecasts. However, this change in the mean can, in some cases, increase bias in the model. Beyond spread, SPP induces systematic changes to the mean state, most notably a general reduction in SIC and a redistribution of ice from the central Arctic toward the ice edge during the winter. These sea ice changes seem to also affect the overlying Arctic atmosphere, producing temperature changes of roughly 0.5°C at 850 hPa, particularly around the Greenland, Barents, and Kara seas. The inclusion of the SPP scheme also results in a substantial increase in the anomaly *correlation coefficient (ACC) for 500 hPa geopotential heights (Z500) over the sub-polar North Atlantic during winter, suggesting better seasonal forecast skill there. The authors hypothesize that this change in ACC arises since the SPP scheme increases sea ice at the ice edge during winter, which increases the meridional temperature gradient and also decreases warm biases near Greenland and the nearby seas. Thus, the correction comes from compensating biases in the model. The authors also demonstrate that sea ice perturbations are considerably more effective at generating ensemble spread in uncoupled ocean forecasts than in fully coupled simulations, a finding with important implications for data assimilation applications.*

Overall Opinion

The study presents in an interesting way of testing ways of correcting the underdispersive nature of sea ice forecasts in numerical forecast models, in this case the IFS. The nature of how changing SIC and SIT impact the overlying atmosphere and aspects of the seasonal forecast is quite interesting and very timely, particularly as the community continues to debate the potential link of Arctic sea ice variability on mid-latitude NH weather variability. The manuscript is written very well, and the explanations are well constructed. The authors did an excellent job as well in presenting hypotheses and then testing them with different experiments, statistical tests, and toy models. These were well-appreciated and should be done more in scientific papers. Well done!

*Overall, the paper is scientifically sound and well-explained. There are some constructive elements of the paper that I think should be addressed to tighten the message of the paper and ultimately improve the presentation and reception of the work. As such, I am recommending **major revisions** before the manuscript is suitable for publication.*

Major Revisions

1. Minimize/Remove Analyses for the Southern Hemisphere (SH). *The main focus of the paper is on the Arctic sea ice and the effects changes in the sea ice forecasts have on NH winter and summer atmospheric variables and seasonal forecasts. The authors also included some similar analyses for the SH sea ice in Supplemental Material and refer to it throughout the paper. While maybe interesting, I found it distracting and overwhelming when reading the paper. The paper has 16 main text figures and 11 Supplementary Figures - that's **a lot** to process as a reader (along with the text). As it is, there is so much to say about the NH to fill this paper that the SH results are not really important. Furthermore, while there appears to be some similarities in the application of the SPP application for SIC in the SH, the authors have not gone further to explore, for example, seasonal forecast impacts in the Southern Hemisphere or other implications for changing the SIC/SIT forecasts. My suggestion is that the authors remove some/all of the figures on SH SIC/SIT and reserve those comments for the discussion and conclusions.*

We appreciate that the paper is lengthy and some trimming would be helpful. We have removed the two figures showing SIT changes in the SH from the Supporting Information (SI). The two figures showing SIC changes are kept, but we have moved them to the end of the SI and only refer to them extremely briefly in the Discussion. We have also tried to trim the paper in a few places elsewhere, and hope these combined changes help make the paper seem more manageable.

2. Z500 Forecast "Skill" Result. *The change in ACC for Z500 in the NH winter is arguably one of the top results in the paper. However, there are a couple of things that I think are important to mention. First, ACC is not "forecast skill" - it is just correlation and measures spatial phase agreement but says nothing about model performance on the field. For example, ACC scores can be high but amplitudes could be off by orders of magnitude in the forecast. Furthermore, the sample size for evaluating the correlation is of concern. So, I would be careful about using the term forecast "skill" in relation to ACC.*

We definitely agree that a high ACC does not mean that the forecast is excellent in all regards. However, we respectfully disagree with the assertion that it is inappropriate to refer to ACC as a measure of forecast skill. It is simply one particular measure of skill, which needs to be viewed alongside other measures (like RMSE) to obtain a complete picture of the forecast quality. We accept that one may question how useful a measure it is, but there is a clear precedent for describing it as a measure of skill in the literature on seasonal forecasts. See e.g. the following examples which include usage from ECMWF, the UK Met Office and NCEP:

https://doi.org/10.5194/gmd-12-1087-2019

https://doi.org/10.1002/2014GL059637

https://www.nature.com/articles/ngeo2824

https://doi.org/10.1175/JCLI-D-12-00823.1

We have therefore continued to refer to it as skill also here. However, we have added a comment emphasising that the amplitudes are low (so the RMSE is high). We have also more strongly emphasised the limitations associated with sample size of 30 when evaluating changes to ACC, noting that Siegert et al. (2017) suggest a sample of 40 is required to have sufficient power for changes to the winter NAO (and hence the same is very likely true for Z500 at gridpoints in the Euro-Atlantic domain more broadly).

Next, while the number of members and simulations may make for a large total sample size, the ***effective degrees of freedom*** *for statistics is an issue because of temporal and spatial correlations in Z500. I'm not familiar with the three-way correlation test used by the authors for checking statistical significance, so maybe serial correlation is already accounted for. For example, improving Z500 forecasts over the North Atlantic will also also change downstream heights over Northern Europe and into Eurasia, so those Z500 changes are not independent* *for testing (Figure 16). The authors should address this and redo their statistical significance to account for the serial correlation.*

Regarding serial correlation in time (“autocorrelation”), note that we are looking at forecasts of DJF Z500, so consecutive samples are separated in time by a full year. Thus the lag-1 autocorrelation is close to zero for CTRL, SPP and ERA5. The effective sample size therefore equals the nominal sample size to good approximation. It also follows that the lag-1 autocorrelation is basically the same in CTRL and SPP (i.e., approximately zero), and so serial correlation in time does not affect the estimation of statistical significance for the difference in ACC between CTRL and SPP.

Regarding correlation in space, we are not making any assumption nor statements about independence of different locations in the domain. Rather, we are testing significance at each grid point and then interpreting the resulting map, an approach which is very common in the literature (e.g. the SEAS5 paper Johnson et al. 2019). We do agree that forecast skill at one gridpoint is not independent from forecast skill at other nearby gridpoints, and hence improvements at one gridpoint are also not independent from improvements at nearby gridpoints. In practice it is extremely challenging to assess the full causal structure of Z500 changes across a broad domain known for high internal variability, and we have not attempted to do so here. However, if SPP has significantly higher skill at two nearby gridpoints, but the improvement at one of them is a causal consequence of improvements at the other, then it does not seem obviously wrong to say that there are significant improvements at both gridpoints. Rather, this simply constitutes extra information which helps explain the differences.

We hope these comments will convince the reviewer that our statistical tests are reasonable as stated. We have emphasised the lack of serial correlation in the revised manuscript.

Regarding the three-way correlation test, the point of this test is that, in our case, we have only modified the forecast by adding SPP. However, the forecast of Z500 at a gridpoint is going to be driven by lots of things in the forecast that have nothing to do with SPP or sea ice more broadly, and thus one expects, a priori, that the two forecasts should be similar to some extent. This is information that one should use to update your null hypothesis, in a Bayesian sense. The test does this by taking into account the joint correlation between the two individual forecasts (SPP and CTRL). If the joint correlation is high, it suggests that the forecast was already strongly constrained by the information contained in the baseline CTRL model, and hence even small changes to skill with SPP could be significant. Conversely, if the joint correlation is low, then larger changes to skill would be required in order to achieve significance.

3. Hypothesis for Z500 Skill Changes. *The authors offer a hypothesis as to why wintertime Z500 forecasts may improve with the SPP scheme – i.e., that cooling in the model over Greenland and the Kara and Barents Seas strengthens the meridional temperature gradient and thus improves Z500 forecasts. They test this mechanism with a bootstrapping test (Figure 16c), but the results are less than convincing. Setting aside the significance issue* *(see previous comment), the correlation is weakly negative (-0.28), which really means that the change in sea ice edge explains ~8% of the variability in Z500 in the subpolar North Atlantic. That leaves 92% of the variance **unexplained.** So, while the hypothesis is interesting, there is a lot of unexplained variance left. I think this needs to be acknowledged and also related to whether or not the improvements in Z500 skill really are meaningful via their proposed mechanism.*

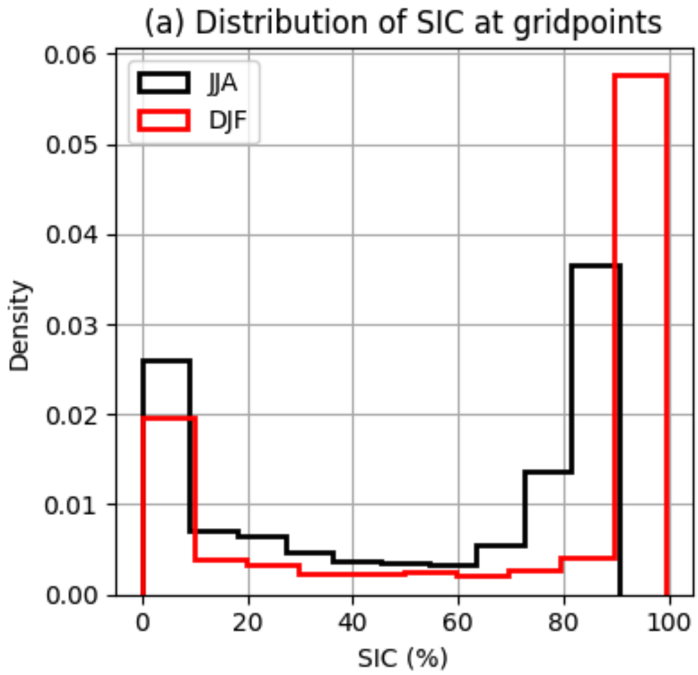
We broadly agree with the sentiment that the explanatory power of our proposed mechanism leaves something to be desired. We think that the proposed mechanism may be part of the puzzle, but is unlikely to be the full story. However, it is a natural hypothesis to make in light of the SPP mean state changes. Furthermore, we think that it's hard to estimate just how much of the signal we are explaining, because some of the changes in ACC are undoubtedly random. In other words, while we explain roughly 8% of the total observed ACC variance, this variance contains physically meaningful variability (i.e., changes causally driven by SPP) and random variability (driven by sampling uncertainty). The proportion of the physically meaningful variability explained by our mechanism could in principle be substantially larger than 8%. Estimating this would however require many more samples (i.e., years) than we have available.

We have revised the presentation and discussion of this hypothesis to emphasise that it's one plausible contributing factor, rather than a conclusive attribution.

4. Summer Results are De-emphasized throughout the Paper. *The authors conduct their SPP experiments on winter and summer conditions in the Arctic, but the authors focus very highly on the winter results. That is somewhat understandable, but I think there are a few more points that could be made about the summertime results. While the impacts of the SIC/SIT changes in summer have minimal effects on the Arctic summertime atmosphere because of how close the atmosphere is to freezing, the authors do not offer alternative mechanisms for why there are significant SIC changes in the summer (see Figure 2). Moreover, the conceptual model in Section 5.2 is less applicable for the summer, but the authors do not offer an alternative mechanism for explaining the summertime results. The authors should conceptualize a model for explaining their summertime results better, as we have observed that Arctic extremes in the summer also impact mid-latitude NH weather patterns.*

We identify two points in your comment. Firstly, a point is raised about the lack of a conceptual explanation for the changes to the SIC mean in JJA. Secondly, a point is raised around atmospheric impacts of the SIC changes in JJA. We address these in turn.

Regarding summer SIC changes, we agree that our view on summer changes was only implicitly evoked in the original draft. In fact we think that the same conceptual explanation applies also for summer, with a small tweak. To state it here clearly, our conceptual explanation for winter changes was (a) the intrinsic pressure for stochastic perturbations to reduce the mean of a strongly bimodal distribution over time (as described by the toy-model demonstrated in Figure 12) and (b) random changes to sea ice motions subject to the constraint that ice convergence in regions with 100% SIC is counteracted by internal ice forces. In winter, this explains why the mean reduces overall, and the ice moves from the centre to the perimeter. What about summer? For (a), note that the distribution of SIC at gridpoints is only somewhat less bimodal in JJA than in DJF. Here's a comparison of the two for the CTRL ensemble:



Thus we think that at least to some extent the same reasoning applies in JJA as for DJF. The one tweak here is that it is now more possible for SIC in the 80-90% category to move up into the 90+% category. This is consistent with the increase in SIC in the 90+% category seen in Figure 11c. For (b), if we compare Figure 5b with the underlying mean state in Figure 2a, we see that the ice is being moved around in a random-looking way across almost the entire domain, with the only restriction being the area of maximum SIC north of Canada. This is a weaker constraint of the motions than in winter, resulting in far more random looking motions.

We emphasise that these two conceptual explanations cannot account for the detailed structure of the change seen in Figure 11c (the "u"-shaped histogram), only the broad strokes (decreased mean + mostly random movement of ice within domain). The more detailed features would likely require understanding the specific physical feedbacks associated with each perturbed parameter, which is beyond the scope of our work.

We have revised the section to hopefully clear all this up, by first presenting the conceptual model in the context of winter and then explaining afterwards the applicability to summer. The above Figure showing the two bimodal distributions has been added as a second panel to Figure 12.

Regarding atmospheric impacts in JJA, Figure 7 shows that the mean atmospheric impact in JJA is very limited in these experiments, which is why we did not look at these further. We agree it's possible that extreme sea ice events could have larger impacts on the atmosphere, but here, we investigate comparatively small changes to mean and variability introduced by the sea ice SPP, for which we do not see robust remote atmospheric impact in JJA. Given your earlier suggestion that the paper is quite long already we hope you will accept this choice. We have added a statement stating that there could be bigger impacts in JJA related to extremes (with references) but that we don't consider this.

Minor Revisions

1. Data Availability. *The data statement is incomplete as presented, as it does not show a repository for the authors' experiments and/or code for analyses.*

The two forecast experiments have now been published with DOI's and so can be accessed externally using the MARS database. Details are given in the revision.

REVIEWER #2

We would like to express our gratitude to the reviewer for a very detailed and thoughtful assessment of our manuscript. We have used many of the suggestions and questions to improve the manuscript, while others are very useful to define future directions of research.

MAJOR COMMENTS

1. Additional information, particularly, whether there are multiple members (an initial spread) of sea ice initial states, plus the ORAS6/ORAS5 citations to provide further understanding [At a minimum Zuo et al., 2017b,a, Browne et al., 2025]. It is difficult to understand the statement (ll. 239-240): "We did not run an uncoupled control (i.e., without sea ice perturbations), as this would be expected to have negligible ensemble spread." My initial reaction was that if you turn off the sea ice perturbations you would have no spread, but then I happened to know that there would be an (small perturbation) ensemble of sea ice initial conditions from ORAS6 to initialized from — others readers may not have this information. Furthermore, providing further background to the possibility of stochastic perturbations to enhance the existing spread of initial conditions (as pointed out in the conclusions/future work) would benefit the motivation of this study.

We have revised Section 2.1 describing the experiments to clarify the ocean/ice initial condition ensemble (there are 11 such initial conditions coming from ORAS6), and provide further references. Unfortunately the various ORAS6 implementation papers have still not been published, but we cite the preprints where available. We also clarify now in this section that SPP-perturbations are not applied when creating the ORAS6 initial conditions, but that the possibility of doing so to boost initial condition spread will be discussed later. We then slightly revised our discussion of this in the Discussion section and Conclusion.

Regarding Section 4.3 on coupled versus uncoupled, we have revised the presentation of the ensemble forecasts being used to remind readers that there is some non-zero spread from initial conditions, but that it will be very small.

2. I would like to be pulled through the explanation of Section 6.2 slightly better, particularly in light of the changes in bias seen with the addition of SPP. The change in mean of DJF Z500 (Figure 6c) and 700hPa winds (Figure 8a) seems to indicate a tendency toward a more negative(ish) NAO. Meanwhile the discussion of Section 5.2 shows that SPP increases the T850 meridional gradient (Figure 16b), which (I believe) is associated with a stronger, more zonal jet stream. This is turn (with less conviction; it is complicated) is associated with a tendency toward a more positive(ish) NAO. None of this is directly responsible for the increased skill seen in correlation (which is independent of the mean state) in DJF Z500 (Figure 16a) and through projection onto the NAO pattern, increased skill in NAO index. However, as the authors and some of their supporting literature have argued, the changes in mean state can modulate (l. 440) the variability – although it is here my knowledge, and therefore my logic may breakdown. Possible graphical improvements to discussion:

- *Include the mean Z500 change (a repeat of Figure 6c) as a third spatial plot in Figure 16 – but this time as a full colour contour, instead of the secondary lines on Figure 6c. Or possibly a no SPPbias versus change in bias Z500 plot as done for T850 in Figure 16b. This allows to see the changes in bias and change in skill side by side. Also a dedicated plot with smaller threshold in lowest (absolute) contour value would perhaps show an anomalous Z500 low over the Ural Mountains (l. 213) that must exist for the anomalous circulation seen previously in Figure 8 (technically, it must appear in a plot of Z700 to give the anomalous U700 fields, but presumably it should be present in Z500, or mean sea level pressure as well).*

Thanks for these suggestions. We agree that a key point was missing from our discussion. We think this was mainly a question of missing text, rather than missing analysis/figures.

The key missing point regards the difference between the Arctic-wide mean Z500 change (a weak NAO- signal) and the local temperature gradient changes (which projects weakly onto NAO+), an apparent "contradiction" you are right to bring up. The overall mean change is a combination of warming in the central Arctic and cooling in the Greenland and Barents-Kara seas. This makes it a priori not completely clear what the Arctic-wide change at Z500 will be, but it turns out to be mostly a positive anomaly across the central Arctic, consistent with the overall tendency for increased SIC there. While this could potentially affect predictability, as Woollings et al. 2023 (https://doi.org/10.5194/wcd-4-61-2023) noted, it is hard to trigger Rossby waves in the central Arctic. It is much easier the further equatorward you go, and in particular it's easier near the ice edge than in the central Arctic. Thus the changes around the ice edge are likely a more potent source of midlatitude predictability, and this is where we see the strengthened gradient (weak NAO+), consistent with the increased SIC there.

Thus we hypothesise that the local change around the ice edge from Greenland to Kara is the most obvious source for the Z500 signal we see, despite the fact that the overall mean state change is weakly in the other direction. This is because (a) the variability is much greater there and (b) it is much easier to trigger waves there (as noted).

We have revised the text to emphasise that we are talking about the local temperature gradient in the subpolar North Atlantic, and that the direction of change here differs from the Arctic-wide one. We also cite Woollings 2023 when justifying the importance of the local change. Given our need to compromise with Reviewer #1 on the length of the paper, we have opted to not expand Figure 16 along the lines you suggest, but have instead added the Z500 bias (and bias change with SPP) to Figure S5 (which previously only showed the T850 bias). We also mention this and added some brief discussion, though we note that the Z500 bias is particularly susceptible to sampling variability, making it less straightforward how to interpret.

3. The skill of the forecast system (Section 6) is evaluated by root mean square error (RMSE), including the spread to RMSE ratio, the bias corrected continuous ranked probability score (CRPS), and finally by correlation. My concern is that the root mean *squared error is not a bias-corrected RMSE (or standard deviation error), while by construct, their CRPS score is. I am not looking for the authors to necessarily change these, but I do need them to justify this accordingly, especially as changes in bias with SPP are definitely present (Section 4). In particular, it is often argued that the spread/error relationship should be between a bias-corrected RMSE and spread – as spread should not be used to correct for bias, as this leads to decreased reliability. Furthermore, the CRPS score is often decomposed into reliability and potential [Hersbach, 2000]. I am somewhat knowledgable, but certainly do not have the access to the experts on the subject that the authors do. I believe that removing the bias before computing the CRPS will mostly (but not exclusively, at least not without additional constraints) affect the potential component. The authors should likely address this in some form. Ultimately, however, this all has an effect on how to interpret the results: If RMSE was calculated after removing the mean bias, not only might it have a better interpretation within the spread/error relationship, but it would also have a near 1-to-1 correspondence with correlation results (reliability). But as it is now, the RMSE results are likely more related to bias. Conversely, the separation between bias and reliability of the CRPS, which is usually handled through the computation of the reliability and potential components, is now only seen through the calculated bias corrected CRPS. Note: Calculation of the CRPS components when validating against an analysis, as has been done here, can be complicated due to the increased memory requirements, so it is not necessarily the route to be taken*

Thanks for this feedback. We agree that the inconsistency introduced by our choices was not ideal. We have revised to now show RMSE, spread/error, and CRPS always using bias-corrected fields. This way the 3 metrics can be compared like-for-like. This does not notably affect interpretation of the results, since our interpretation already included a discussion on the role of biases. We have revised the text accordingly, and hope it reads better now.

4. Section 5: I am in total agreement that the changes in mean state with respect to changes in spread and the introduction of stochastics is a consequence of the bounded nature of sea ice and stochastics pushing the mean state away from the bounds due to the 1-direction limitations at the bounds. Indeed I was convinced at its first mention at the beginning of Section 4. Nevertheless, I applaud the authors for a detailed analysis, including toy models to demonstrate this important point. However, I would suggest one further addition to strengthen the point, and allow me to suggest a possible 2nd (secondary) mechanism due to irreversible dynamics (like ridging).

> *(a) The thickness plots (Figures 3 & 4), seem to indicate an increase of thickness along coastal northern boundaries of the Canadian Archipelago and Greenland, along with the east Siberian coast, particularly in the winter (Figure 3b). This I suggest might be down to the stochastics leading to not easily reversible processes such as ridging, that could lead to a build up of thickness, especially in areas where ice is constrained by physical boundaries. Note: Dynamic processes would not necessarily change the grid cell average thickness, as it will just redistribute the ice within the grid cell, however, the accompanying decrease in ice concentration, would then allow for more ice to be advected in and actually increase the mean grid cell thickness. This could account for the increase in mean thickness in these areas, along with some of the decreased sea ice concentration. The decrease of sea ice thickness near the thickest ice right along the north coast of Greenland might then*

indicate a threshold where the stochastics does not favour increased ridging, but enough disturbance to the state to facilitate its breakdown. But also note: I am not a sea ice dynamicist.
Aside: It is never mentioned that the sea ice thickness used in your analysis is the grid cell mean thickness, presuming that is the SI3 model variable. I find it useful to always remind readers of this.

We thank the reviewer for suggesting this mechanism. We have adopted their suggestion. In the revised manuscript, we offer asymmetric impacts from symmetric perturbations to irreversible processes such as ridging as a likely explanation for the increase in sea ice thickness in regions where ice typically converges. We have also clarified which definition of sea ice thickness we employed throughout the manuscript.

(b) An additional simple piece of information I would think would further confirm some of the manuscripts suggestions would be to tabulate (they are single numbers – unless you produce a yearly timeseries) the hemispheric change in Sea Ice Extent, Sea Ice Area and Sea Ice Volume with and without sea ice SPP. The sea ice extent should increase with SPP, as the decreasing high concentration ice does not affect the sea ice edge, but increasing the low concentration sea ice should increase the sea ice extent. If the stochastics is essentially working on free ice, these changes in concentration at either end may spread out the ice edge, but leave the sea ice area the same. Finally, if sea ice dynamical processes start to be important, the sea ice area might change, but the sea ice volume would (maybe) be left unchanged (just re-distributed). This likely might only work in the winter, as summer time thermodyanamics, like sea ice being ejected into warmer waters, is likely to affect results. Nevertheless, this would likely be a simple useful diagnostic to have, which is not necessarily easy to ascertain from just viewing changes in the spatial distribution of sea ice concentration and thickness.

These are interesting suggestions. We have checked that, in DJF, the sea ice extent does indeed increase slightly with SPP (by around 0.2%), the sea ice area goes down (by around 0.2%) and the sea ice volume goes up by around 2%. The slight increase in sea ice extent area could indeed be due to the effect that the reviewer refers to and that we also investigate with our simplified model. The increase in sea ice volume might be due to the previously discussed non-symmetric effects of perturbations on some parameterisations like the ridging scheme. Further sensitivity experiments perturbing only a single parameter would be needed to investigate which parameter perturbation leads to which mean state change. We note that there is a lot of regional and seasonal heterogeneity in the SPP impact. As discussed elsewhere, the regional and seasonal response of the sea ice to SPP is markedly different for each parameter perturbation, so the resulting changes to the mean state are likely a residual of regionally and seasonally differing single-parameter contributions that partly cancel out.

We now state the 0.2% increase in sea ice extent and 2% increase in volume in DJF in Section 4.1 (on the basic impact of SPP on sea ice), and allude to these again in

Section 5.2. We have opted to not add a more comprehensive Table, because of our concerns about the length of the manuscript raised by Reviewer #1.

5. Figure 10 (and discussion): I mostly agree with everything discussed with regards to Figure 10. However, I have a few comments.

• Scatter plots are easy and convenient, but they have a tendency to sometimes give false impressions due to the overlaying of points on the plot. In this context, I would think a 2-dimensional probability distribution (or 2-D binned plot) would be better. I.e. Bin the number of occurrences as a function of spread and mean change, and then plot frequency of (normalized) occurence as colour or gray shading on 2-D plot. Instead of repeating the full (black dots) scatter plots for rows 2 and 3, you would then simply replace them with 2-D PDF's for > 80% and < 20% concentration.

• This would negate the need of the 90% of points contour. Note: A contour encircling 90% of points is not unique (otherwise it would not be possible to gerrymander election districts). Enclosing 90% of points using the smallest area, however, is unique – and I suspect what is shown.

• The scatter plot for JJA is not more Gaussian (ll. 303-304: Actual statement "this non-Gaussianity is less pronounced"), as it is not centered around the zero mean change. Rather it is missing the positive (blue) node of the bi-modal distribution due to thermodynamics reducing low concentrations that may have increased due to stochastics. I.e. Thermodynamics suppresses this mode in the summer – and likely(?) elongates this mode in the winter, as thermodynamics further increases the low concentrations increased by stochastics. More bluntly, all points get (on average) pulled to the left (lower concentrations) in JJA and to the right (higher concentrations) in DJF due to thermodynamics

Apologies, the text in the manuscript was unclear. In fact we do exactly estimate the 2D distribution using a kernel density estimator, and the contour "enclosing 90% of the points" is actually the 90th percentile based on this estimate. Thus the contour encloses the 90% densest points. We have revised the text to clarify this. Regarding your suggestion of how to revise the figure, we feel that omitting the scatter points altogether from the middle and bottom rows makes it harder to interpret, because we want to allow readers to see clearly which sea ice regimes are responsible for the non-Gaussian structure in the top row. This is in our view most easily done by highlighting the actual points. However, we have altered the middle and bottom rows in the following way: the red contour now shows the 90th percentile of the estimated 2D PDF *based on the subset of points being considered*. So for Figures 10c and 10d, we estimate the PDF of the mauve points, and in Figures 10e and 10f we estimate the PDF of the blue points. This way one can get a sense of where the majority of the points are actually located (inside the red contours) while still allowing an easy comparison with the top row. We hope you will find this compromise reasonable.

Your final point is a good one, and we agree that is likely what is going on there. We have added a brief statement stating this.

MINOR COMMENTS

1. 7: "(enhance) by around 10% relative to an unperturbed forecast" This is somewhat confusing and harks back to Major Comment #1. A reader reading "unperturbed" forecast will naturally(?) assume a zero-spread (deterministic) forecast and wonder how you can increase zero by 10%. I'm afraid you may have to expand the explanation somewhat ("relative to a forecast where pertubations are not applied directly to the sea ice").

We rephrased to "relative to a forecast without SPP".

2. l. 18: "tendency perturbations" Although I have no desire for additional acronyms to be added unnecessarily, I wonder if it would not be advantageous to somehow mention this is what is commonly referred to as Stochastic Perturbed Parametrization Tendencies (SPPT)

We added in "SPPT" with the spelled out acronym.

3. l. 79: Should mention what NEMO4/SI3 is an upgrade from (I believe NEMO3.4/LIM2 based on ORAS5 documentation).

We added this detail.

4. l. 81. Technically the 5 category ice is not a property of SI3, but an adjustable option (i.e. could equally have chosen 1 category or 10 category). I do not see any perturbed sea ice parameter other than snow conductivity that might be affected by the number of thermodynamic layers, but it might be useful to document the number of ice and snow thermodynamic layers as well (ORAS6 has 4 ice and 1 snow thermodynamic layers, which I suspect would be the same here), along with whether a melt pond scheme has been activated (yes, presumably if you perturb the meltpond parameter rn pnd flush).

We added the additional details on the thickness categories (5 ice thickness categories each with 4 thermodynamic layers, and a single snow layer on top). The inclusion of the melt pond scheme was already mentioned.

5. ll. 83–86. I would think the properties of the sea ice grid would be far more important to discuss than the atmospheric grid. Beyond mentioning that the ocean grid is 1/4◦, I think it would be wise to mention it is (based on Browne et al. [2025]) the eORCA025 tripolar grid, with grid sizes between 3 and 12 km [Subich et al., 2020] in the Arctic, and down to 4 km [Smith et al., 2021] in the Antarctic. This would be of importance to the covariance length scales generated by the Shapiro Storto and Andriopoulos [2021] or Perlin Noise Perlin [2002] filters.

We state now the grid used and that the average grid box size in the polar regions is around 15 km with some grid cells as small as 3km.

6. ll. 95-97: The SMMR/SMMI/SMMIS OSI-450-a products the authors use for validation may easily vary from other sea ice concentration products by (approximately) amounts near to the changes in spread seen in this manuscript [Renfrew et al., 2021, Peterson et al., 2022, Niraula, 2023], at least in the marginal ice zone. As SEA6 would be initialized by this product, it would remain the best product by which to estimate forecast error. However, it might be a good idea to state that there is some uncertainty in the sea ice concentration "observations."

We thank the reviewer for this important comment. We would like to point out that - while changes in the spread in SIC are small, the overconfidence in the baseline ensemble forecasts is not, so we do not expect our conclusions to be significantly affected by observational uncertainty. In the revised manuscript, we added a comment acknowledging the uncertainty of observations in SIC.

7. ll. 117-118. You might want to mention you only used correlation in the context of verification of Z500?
[I kept expecting correlation results that never came.]

Done.

8. Section 2.2: I will have to admit to not knowing Weather and Climate Dynamics requirements. Is a open code respository of analysis tools required for publication.

Data needs to be open, code is optional. Much analysis here was based on internal ECMWF tools and so cannot be immediately shared. However, we have made the data publicly available (see Data Availability section), and ultimately we are mostly just computing simple ensemble statistics (mean, spread, RMSE and correlation) so we hope this will allow for sufficient reproducibility.

9. ll. 132-135: The stochastic perturbation time scale (1 day) is given, but the covariance length scale (number of iterations of Shapiro, and presumably also the Perlin Noise filter) is not given. The actual length scale would be approximate and geo-spatially dependent owing to the grid cell variation in the eORCA025 grid, particularly in the Arctic, but an approximate range could still be provided.

Apologies, we had accidentally omitted this key piece of information. We now clarify that we use 250 iterations, corresponding to a length scale of around 750km. We have not included further information about Perlin Noise in this section, but have added a brief comment in Section 4.3 (on coupled vs uncoupled simulations), stating that the main difference between the Perlin Noise fields and the iterated Shapiro filters is a somewhat larger length scale with Perlin. This is because Section 4.3 is the only place that Perlin noise is used. The fact that there is this one discrepancy here with other experiments is unfortunate and adds some extra complexity to the presentation one way or another, but on balance we felt it's better to try to keep this complexity contained within Section 4.3.

10. l. 168. I suspect this statement is relatively independent of the definition of ice edge, but please define the definition of ice edge (sea ice concentration = 0.15?)

Thanks for catching this, we now include the definition, which is indeed as you guessed. We did not test sensitivity of this statement to the exact choice of threshold, but we do not expect there to be much sensitivity. This is because the SIC gradient near the ice edge is often very steep, and on the 1x1 degree evaluation grid, changing the threshold a bit would often not change the ice edge latitude diagnosed in CTRL at all.

11. Figure. 1-4: Fram Strait seems to be a relatively active area (likely to always be an active area) in changes to mean state and spread of both sea ice concentration and thickness. Were any calculations performed to diagnose changes to mean and spread of sea ice transport out of the Arctic choke points (i.e. besides Fram, Bering/Davis/others [see e.g. Smith et al., 2024]) – or even just Fram Strait? This would seem to have implications for Section 5.

These are interesting suggestions, but we have not performed such computations. However, we have submitted a follow-up paper on the impact of SPP in sub-seasonal forecasts, which does look at one such chokepoint as a case-study.

12. Figure 6, 7 & 17 (others?) It is not indicated whether the solid or dashed lines correspond to positive contour levels. I assume it is the solid lines – but that is an assumption without context in the caption.

Oops, thanks for catching this. We now state this.

13. Figures 6 & 7: This is the first and only time that the change in spread is expressed as a percentage change. You explicitly write this out in the "units" of the colour bar, so as to differentiate with the usage of % for concentration and concentration changes (good). But perhaps this should also be expressed in the caption explanation for Figure 6b – which now only indicates it is a change in spread (i.e. not a fractional/percentage change in spread relative to the control spread – I assume).

Thanks for catching this mismatch. It's the caption that's wrong. We've corrected this.

14. ll. 239-240: This is were I missed an explanation of whether there are perturbations or not in the initial conditions. If there are no perturbations to the initial conditions there will not be negligible spread, there would be no spread. If there were perturbations to the initial conditions, then there should be spread, but not necessarily negligible spread — unless the spread in the initial conditions is quite undispersive.

We now remind readers that there is spread in the initial conditions of the ocean/ice (11 distinct initial conditions), so there is some ensemble spread even in the ocean-only simulations. This spread is not zero, but we think likely to be small enough to count as negligible, simply because the spread in the initial conditions is indeed particularly underdispersive, as mentioned in the Introduction. Indeed, using SPP to help with this is an active area of work at ECMWF. We point these things out in the revision.

15. ll. 241–244: Perlin Noise, Does this change the characteristics of the covariance lengthscale in any appreciable fashion?

Good question. The answer depends on the free parameters of the Perlin noise algorithm, which allows you to tweak the lengthscale. For these experiments the lengthscale was closer to 1000km, as compared to the 750km lengthscale with the Shapiro noise configuration we used. We added a very brief comment on this. Our experiments suggest that the impact of SPP on the sea ice is largely independent of the details of the noise field, and in particular Perlin noise and Shapiro noise produce more or less the same impact despite these differing lengthscales.

16. Figure 9b-c: The spatial plots are quite obviously masked to correspond to the hemispheric mean in Figure 9a. While I do not think it is necessary to do this masking, if done, if should probably be documented in the caption.

Thanks, you're completely right that we failed to specify this. The fact that gridpoints with no ice have been masked is now stated in the caption.

17. Figure 9b: Note: The month-to-month temporal variability is essentially (or at least strongly dominated by) the seasonal variability. I.e. You are seeing where the sea ice changes most between summer and winter – and not month to month timescale variability as you might see in atmospheric fields such as the Z500.

Yes, we agree. However, we expect that the *difference* between the two simulations is dominated by changes in the genuine month to month variability, as opposed to changes in the seasonal cycle itself, since the sea ice cycle is strongly tied to the solar cycle which is the same in both.

18. ll. 249–255: An observation. I am not particularly surprised independent perturbations do not lead to as much increased spread as in the uncoupled simulations. The increased spread in the uncoupled simulations is presumably coming from non-linear or non-orthogonal interactions between the perturbed fields. My impression is that the coupling with the atmosphere would more than adequately generate that. Plus, I would assume that the atmospheric perturbations are not provided by the same perturbation fields either, nor with the same temporal and lengthscale coveraiance characteristics, so there is already an existing non-orthogonal independence built in.

It wasn't so obvious to us in advance, but in hindsight we agree that the result seems plausible or even unsurprising. We hope that readers will still find value in these experiments providing "intermediate" steps to confirm where the different response in coupled and uncoupled simulations is coming from.

19. Figure 11 (JJA): I seem to recall that the 60% sea ice concentration is important for air/sea flux exchange. Is the peak in JJA spread and mean change (peak in amplitude of change, not neccesarily frequency) possibly related to thermodynamic considerations?

This is a great comment, and we think indeed there may be important thermodynamic effects explaining the shape of this histogram more broadly. The reviewer is probably referring to the fact that at these intermediate concentrations, the sea ice surface roughness that the IFS sees has a maximum. While we'd love to investigate these effects more, we think we need to relegate this for future work.

20. ll. 390–391: "At the coastlines, where validation is more robust, there is a mixed impact." I am not sure I completely agree with this statement. Firstly, do you mean at the coastlines, or at the sea ice edge / marginal ice zone? In both cases, however, the SMMR/SMMI/SSMIS satellite product has a relatively large footprint that gives the observational product a rather large uncertainty in these regions. Indeed, the OSI-SAF product has been shown to have a large degree of uncertainty in the marginal ice zone [Renfrew et al., 2021].

Thanks for challenging this. We agree, and have removed the assertion that SIC observations are more robust around the coast or ice edge.

21. Most ensemble verification of sea ice is done bypassing complete knowledge of the sea ice concentration, and focusing on more important changes mostly to the marginal sea ice by evaluating skill on the predictability of the probability of ice (fraction of ensemble members that exceeds some sea ice concentration threshold, most commonly 15%) as in Zampieri et al. [2018], Peterson et al. [2022] and others. I would be curious if any of the conclusions regarding sea ice predictability would differ from taking this standpoint.

We decided to focus on sea ice area here because we want to study atmospheric impacts. Focusing more on sea ice extent and the ice edge gives results that are qualitatively consistent with our findings. We will address this in the follow-up study that we previously mentioned.

22. There is no thickness assimilation in the initial sea ice state, so there are likely significant biases in the sea ice thickness. It would be an interesting topic (not necessarily for this manuscript) to evaluate whether introducing stochastic perturbations to the sea ice increases or decreases these biases – and if then actually assimilated increases/decreases the ability to properly forecast ice thickness evolution over a season. The Copernicus combined Cryosat/SMOS reprocessed thickness dataset is available from 2010 onward (from 1995, but I would not necessarily recommend the pre-CryoSat2 period) which should give a suitable window to produce (DJF) sea ice thickness forecast errors (https://data.marine.copernicus.eu/product/SEAICE_GLO_PHY_CLIMATE_L3_MY_011_013/description.

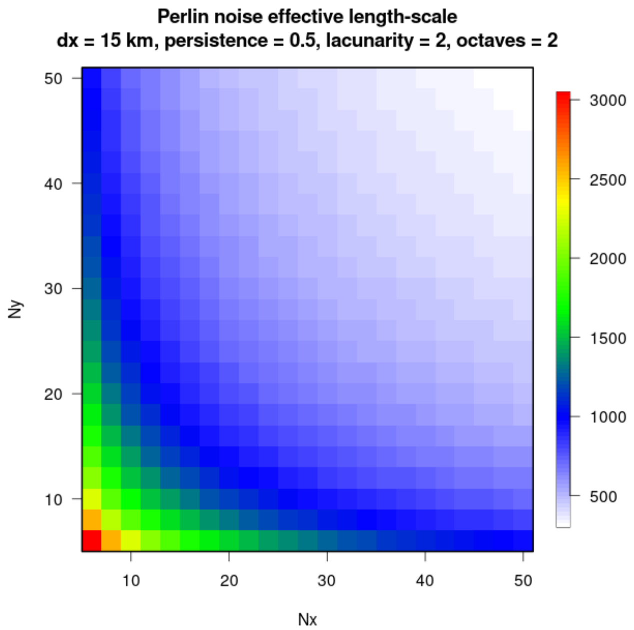
Thanks for this interesting suggestion! Indeed, there is ongoing work at ECMWF towards assimilating sea ice thickness, as well as using SPP for data assimilation. We have passed this comment on to the relevant colleagues.

23. There is no mention of any effects on the sub-surface ocean in the manuscript. While there is unlikely enough sub-surface observations to meaningfully test the changes to the sub-surface forecast skill, it might have been interesting to at least show pan-Arctic average changes to both spread and mean T/S profiles. This could at least give a broad view of how deep the stochastics might affect the sub-surface solution.

We agree this is interesting to consider, but we are concerned about the length of the manuscript already so have opted to not add such additional analysis. However, you may be interested to know that we are currently exploring similar stochastic perturbations in the ocean (i.e. in NEMO itself, rather than SI3), and there we are looking at depth as well. We expect this analysis to be published in due course.

24. The Perlin [2002] paper was not easy to find. Please provide the DOI: https://doi.org/10.1145/566570.566636. Also, I believe "noise" should be capitalized as "Noise." It is not generic noise — but a proper name "Noise algorithm." That being said, I did not find the given citation particularly useful for the context I was looking to answer: What covariance lengthscales are generated by (iterations of) the Noise algorithm

Our apologies, we have now added the DOI and capitalised Noise. The length scale of the Perlin Noise is a function of 5 free parameters: persistence, lacunarity, and octaves, and the size of the sub-domains used when generating noise (given by Nx and Ny parameters). In addition, it is scaled by the basic size dx of the gridbox the noise is acting on. In our case dx is approximately 15 km (the rough size of gridboxes in the polar regions). Here is a figure showing the length scale in this context, courtesy of Andrea Storto:



We have included a brief description of Perlin Noise in Section 3 (describing the SPP scheme).

Centre is mis-typed "cdntre" (l. 55). I am surprised that one got past internal review!

Corrected.

Figure 3: Not an error — just an observation: The sideways unit m is easily misinterpreted as an upward oriented E. I was originally thinking that units were neglected on the colour bar – and wasn't sure what E stood for, until I realized the colour bar title (units) was rotated. I would suspect any remedy might not improve the possibility for confusion (and it could only be me than is confused), so I will not suggest or insist on any.

We think the obvious way to clarify this would be to spell out "metres", but we know from experience that at type setting this will be reverted back to "m" anyway so we have left it as is.

l. 236: Considering the subject matter, I would consider writing ocean and sea ice only forecast.

We revised to say "ocean-and-sea-ice-only forecast".

Figure 19: My eye is drawn to a dipole which somewhat unnaturally seems to be split across the 60◦E/120◦W meridian, at least north of Svalbard and then south again to the Canadian Archipelago. I do not see this in any of the other plots, but given that this is a comparison between a coupled and uncoupled simulation, I wonder if this is the only plot based upon the output from NEMO/SI3, with all other plots being based upon sea ice output from (and interpolated to) the atmospheric grid. The default assumption must be that this is real, but I do want to beware of possible plotting artifacts. The most obvious candidate might be the north fold in the ORCA025 grid (and I have produced many plots, some published with discontinuities across that line), but that is not along that meridian — as the north fold is along the 73◦E / 107◦W meridian. Nevertheless, I just want to check whether there might be some other possible plotting artifact.

We presume you meant Figure 9 here. Thanks for raising this potential pitfall. We have double-checked our scripts and we do not think there was a mistake when interpolating the fields.

l. 335: 9/nine. I do not normally pay attention to this, but this may be an instance where typographical conventions would suggest that 9 be spelled out as nine.

This will be decided by the type setters at the end, so we will leave it and let them edit.

l. 558. I believe the reference type should be [Hersbach et al., 2023] and not Hersbach et al. [2023].[\citep and not \cite if using LATEX].

Thanks, we fixed this.

ll. 575, 577, 586, 589, 591, 602, 608, 615, 626, 630, 642, 644, 646, 650, 656, 663, 665, 673, 681, and 697.There are extra https://doi.org in the DOI of many references — and clicking on the front part will take you to non-existing pages.

Thanks, this was related to a Bibtex mismatch. We have fixed these now.