RESPONSE TO REVIEWS OF WCD PAPER (2023):

“Predictable Decadal Forcing of the North Atlantic Jet Stream by Sub-Polar North Atlantic Sea Surface Temperatures”

RESPONSE TO REVIEWER #1

GENERAL COMMENTS FOR BOTH REVIEWERS
We thank the reviewer for their thoughtful comments and feedback. Based on these comments and those of the other reviewer, we have carried out considerable revisions. We provide a summary of the biggest changes:

- Several sections have been shortened or cut entirely (such as the section on stochastic atmospheric forcing) in an effort to make the paper more concise. The number of figures has also been reduced. Note that Section 6 was not cut entirely, as the second reviewer suggested. The analysis of aerosols and the AMOC have been retained in the revised manuscript, since we felt these added value to the paper (see responses to 2nd reviewer).

  We welcome further feedback from both reviewers (and the editor if they are reading) on the value of the trimmed down Section 6.

- The discussion has been largely removed from the various results sections and is now relegated to its own section close to the end.

- The statistical significance testing has been modified somewhat, so that we now consistently adopt a null hypothesis where atmospheric timeseries are modelled as white noise and SSTs using the Fourier phase shuffle method. Details of the tests and thresholds used are now more consistently made clear, including in the captions of figures.

- Appendices are now in a separate Supporting Information file.

- Analysis of heatflux variability, such as gridpoint correlations between SSTs and heatfluxes, have been removed. Based on comments from Reviewer #2 and discussion with co-authors and other colleagues, we now feel that much more careful analysis would be required to compare the heatflux variability across the different datasets we use.

  The related analysis where we compute correlations between daily SSTs and surface temperatures in the SPNA region has been retained, as we judged

this was less prone to the subtleties associated with the heatflux variability. The speculation on the role of roughness lengths has been kept, but moved to its own section in the Discussion and is phrased more cautiously.

We also need to flag two errors in the original draft.

Firstly, there was an error in the original Table 2 (correlations between SPNA and JetSpeed). The values for ERA20C and ASF20C reported were based on a slightly different SST domain corresponding to the Labrador sea, due to the fact that much earlier versions of this work focused on that region. The corrected correlations using the SPNA domain are now reported: the reviewers will note that they are very similar to the previous values, and that questions of statistical significance are not altered (which is why we failed to notice before).

Secondly, the original draft suggested that detrending of NAO and jet speed/latitude timeseries was carried out, but this was incorrect. We had tested both with and without during the initial analysis and found there was no meaningful impact on the choice to detrend or not, due to the fact that the trends were very small (and probably just reflecting internal variability). As an example, the correlation between non-detrended 30-year running means of the SPNA and JetSpeed in ERA20C is 0.97, while using detrended timeseries gives you 0.92, with both being significant. The story is similar with other quantities: questions of significance are not altered. It's not entirely clear what the 'correct' thing to do is, but we ultimately opted to not detrend for simplicity. The fact that the manuscript suggested we detrended anyway was due to some copy-and-pasting from older manuscripts of the lead author that we failed to tweak. This is now properly clarified and corrected in the revised.

Apologies for both of these!

All the revised parts are highlighted in RED in the revised manuscript.

We now respond to each comment in turn.

Major comments:

Reviewer #1: *Line 89: The authors argue that the use of the atmosphere-only ASF20C allows to isolate the atmospheric response to ocean variability. However, prescribed SSTs could still be affected in part by 'real' atmospheric forcing, as they come from observations?*

We are guessing that the reviewer meant Line 79 here.

Yes, that's true when looking at the atmospheric variability across multiple start-dates. However, within a given ASF20C forecast (i.e. the forecast corresponding to a particular winter) the atmosphere is purely responding to the ocean and not vice versa, so that on seasonal time-scales ASF20C does allow for

such an isolation. This point is explained later in the paper (L245 onwards in the revised draft), and is touched upon again in the new section on causality (paragraph around L565 in the revised).

Put differently, we are not claiming that the prescribed SSTs are not carrying some imprint of atmospheric forcing (surely they do), but that to the extent there is any interaction between the SSTs and atmosphere in the ASF20C forecasts it can only be forcing from the SSTs and not vice versa. This does not solve the issue of causality, but it does seem to offer some simplifications.

Reviewer #1: *Lines 202-203: Recent evidence suggests decadal predictions failed to capture the latest (2010s) trends in the NAO and jet variability. This would suggest predictability can change so how reliably can you assimilate ASF20C/CSF20C to genuine decadal forecasts?*

We agree that it's not clear that the skill seen in 1900-2010 can be expected to continue indefinitely into the future. We only make the claim that in this period, it appears that ASF20C/CSF20C behave like the DPLE decadal forecast, and that in this period the skill seems to come from the SPNA. We have removed suggestions in the Conclusion about the potential for future skill to avoid in response to this.

We would add that in the context of a decadal forecast spanning 1900-2020, the 10-year period 2010-2020 is basically just 1 datapoint, so it seems hard to rule out that the forecast models just had a bit of bad luck the last decade. We note that the correlation we found for 10-year means was 0.64 for DPLE, which only explains 41% of the variance, so bad 10-year periods are to be expected.

Reviewer #1: *Line 242-243: I would include here a brief discussion of the alternative explanations, or a least providing a list of examples, so that it is easier to reconnect with Discussion and Conclusion.*

We have rephrased this, because we actually only came up with 1 obvious alternative explanation, namely that it may be that ASF20C/CSF20C obtain most of their 'skill' from the initialisation, while DPLE obtains most of its skill from representing slower frequency processes better. We now mention this straight away following these lines, and again in the conclusions.

Reviewer #1: *Line 258: Are you referring to the midlatitudes here? A reference would help here, as you later use this relatively low decorrelation value to exclude the troposphere as a source of predictability.*

Yes, sorry, we meant the midlatitudes. We have added a recent reference for this.

Reviewer #1: *Line 264-265: Can you add a few words on what is the visible impact of stratospheric bias shown by O'Reilly et al. (2019)?*

We have done so. Basically, the QBO is substantially weaker in ERA20C than in reanalysis which assimilate the atmosphere (like ERA-Interim/ERA5), the downward propagation is too small, and the association with the NAO on seasonal timescales is non-existent (lines 510-516 in the revised).

Reviewer #1: *Lines 300-304: On Line 254 you said that weak signals could be due to deficiencies in atmosphere—ocean coupling. Would it not be the case that DPLE has an advantage on ASF20C which might compensate its smaller size, to some extent?*

We agree that such a compensation could occur. However, CSF20C is also coupled, has more ensemble members and twice the sample size to DPLE, and also highlights the SPNA. So even if one biases somewhat towards coupled forecasts, it seems fair to weight CSF20C above DPLE.

Reviewer #1: *Lines 347-351: Wouldn't the SSTs that are fed to ASF20C implicitly feature the full time variability, if the come from observations? If the atmosphere is well simulated so would be its coupling to SSTs, so you might be including some oceanic response in your analysis.*

The covariance between SSTs and jetspeed on interannual timescales in ASF20C is the sum over products SST_n*JetSpeed_n, where n runs over the DJF of each year 1900-2010. This covariance being large therefore depends on these products being large. Each such product depends only on what ASF20C is doing within a single season (a single value of n). SST_n is just prescribed, while JetSpeed_n is simulated by ASF20C.

Therefore, the only way ASF20C can produce significant interannual correlations is by correctly simulating what JetSpeed_n is going to be. For each DJF (each value of n), this simulation depends only on the initial conditions and boundary forcings of that given DJF, since these are the only data that ASF20C knows about. Thus, ASF20C can only produce significant interannual correlations between SST's and JetSpeed by simulating how the JetSpeed is forced by the initial conditions/boundary forcings (including SSTs). In other words, if there is any causal interaction between SST_n and JetSpeed_n in ASF20C, it is necessarily from SST's to JetSpeed.

There is one caveat to this though, which we hadn't explained properly in the draft, and which the reviewer might be thinking of here as well, namely the possible existence of a confounding driver. If JetSpeed_n is simulated well due to some other source of skill, and the observed SST's are responding to the jet in the real world, then this would induce correlations between JetSpeed and SSTs in ASF20C which are not a result of a causal link from the SSTs.

Our assertion that interannual correlations necessarily imply causality from the SSTs is therefore only accurate if one assumes that the skill is definitely coming from the SSTs. This is of course what we have been assuming in the paper, but it might be wrong!

Upon reviewing the paper we generally felt that these points, and other things around causality, were not that well explained or covered, including this point raised by the reviewer. We have therefore decided to collect all the different discussions of causality into its own subsection in the new Discussion. We hope this will help make the discussion more readable.

Reviewer #1: *Lines 352-355: I found this paragraph somewhat hard to follow, mostly because you refer to later sections, which makes the reader go back and forth. I think it would help if you elaborated a bit more on the argument presented here (when you say negative fluxes, are you referring to anomalies or absolute values? Negative anomalies could still be associated with heat flux from the ocean into the atmosphere, perhaps simply anomalously weak).*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #1: *Lines 492-494: Colder SSTs lead to lower moisture availability, which can weaken eddies and consequently the jet, but this is not accounted for by dry models. Recent (and less recent) studies have indeed suggested that the representation of precipitation can be behind the current bias in model representation of storm tracks (too equatorward, too zonal; see Willison et al., 2013; Schemm 2023; Fuchs et al., 2023), consistent with the importance of the role played by latent heating in baroclinicity restoration (Papritz and Spengler, 2015). Could you comment on this and possibly add some discussion in the text?*

We have added a caveat about biases in dry models, as suggested. We note that the results of Baker et al. are generally consistent with the findings of Baker et al (2019) using a full moist GCM.

Reviewer #1: *Line 542: Here and possibly elsewhere, can you specify what convention you are using with regard to the sign of surface heat fluxes? The ECMWF convention of positive (negative) fluxes into (out of) the ocean, or the opposite? I would also specify this in the caption to Figure 13.*

This is no longer relevant due to the removal of heatflux analysis.

Minor comments:

Reviewer #1: *Lines 135-140: I would move these lines to the start of the DPLE subsection, so as to make clear from the start what model the 40 members come from.*

We made the suggested change.

Reviewer #1: *Line 166: Why do you remove a linear trend from the NAO? Not arguing against it, but a few words behind this choice would help.*

See the “general response to both reviewers”.

Reviewer #1: *Line 227: I would move the comma to after ‘identical’ (…very close, but not identical, to the…)*

Thanks, yes we agree and made the change.

Reviewer #1: *Line 311: I would remove ‘again’, as you have not really discussed it until here.*

Done.

Reviewer #1: *Line 312: Can you add a few words on what ‘false discovery rates’ are?*

We now write: “It is worth mentioning the phenomenon of false discovery rates, namely the fact that if one looks for correlations across sufficiently many predictors then some are bound to be significant by chance.”

Reviewer #1: *Lines 326-327: Confidence intervals for ERA20C, ASF20C and CSF20C are not shown explicitly? The values in Table 1 are related to the average? Not asking you to show this, but perhaps add ‘not shown’ and clarify how the confidence intervals are obtained in the caption to Table 1.*

The original text wrote:

“Note that the confidence intervals for ERA20C, ASF20C and CSF20C are almost identical, so we use their average intervals as a common confidence interval for all three.”

In other words, the confidence intervals given in Table 1 for these datasets is the average of all 3 individual confidence intervals. To hopefully make this clearer, we have rephrased to:

“Note that the confidence intervals for ERA20C, ASF20C and CSF20C are almost identical, so we only report a single confidence interval encompassing all three, obtained by taking the average across the three individual intervals.”

Reviewer #1: *Line 340: Add ‘as’ or ‘which is’ before discussed.*

We added an ‘as’.

Reviewer #1: *Line 392: Never came across the word ‘straddle’, one never stops learning! Would you consider an alternative using more common words?*

Haha, ok! We replaced “straddles” with “is very close to”.

Reviewer #1: *Line 455: I would put ‘similarly for warm anomalies’ between parenthesis rather than in a new sentence. Also, anomalously cold SSTs … cool rather than cools.*

Both changes were made.

Reviewer #1: *Line 465: To improve reading, I would add a comma after ‘to begin with’. On the same sentence, isn’t ‘localised with’ more natural in English than ‘localised to’?*

The phrasing “localised to” seemed most natural to a (small) random selection of British colleagues surveyed, so we’ve left that as is. The comma has been added.

Reviewer #1: *Lines 465-467: I would refer to panels of Figure 9 more specifically. Line 465 seems to refer to Figure 9a, while the following line to Figure 9b, is that correct?*

Figure 9 isn’t meant to exemplify the evolution being described: the initial baroclinic stage would be taking place within 10-14 days, while 9a shows November means (i.e. 30 day average). Figure 9 is supposed to just show that after some time has passed the structure is roughly barotropic. We have rephrased to: “Figure X shows that in DJF, the structures are relatively barotropic in nature, consistent with this tropospheric pathway having taken place.” We also added a line stating that the same conclusion is drawn if regression coefficients are plotted instead (in response to a question by one of the authors).

Reviewer #1: *Line 470: I would add a boxes in Figure 8b to help locate the two regions you define.*

We made the suggested change.

Reviewer #1: *Line 480: Can you add units of measure and use scientific notation for large numbers? E.g., 1x10^-4 for f and 2.553 x 10^6 m for dy. Also, I would use the equal sign instead of the approximation, given that those are the exact values you have used in the equation to calculate the time series. On the same line, 'here we have assumed' sounds better to me (up to you though).*

We made all the suggested changes.

Reviewer #1: *Line 526: I would either move 'in ERA20C,' at the beginning of the sentence or add another comma before 'in', it would read better.*

We moved 'in ERA20C' to the beginning of the sentence.

Reviewer #1: *Lines 530-531: Can you briefly discuss the meaning of negative correlations?*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #1: *Line 549: Similarly to an earlier point, I would add a comma before 'in ASF20C,'.*

We added the comma.

Reviewer #1: *Lines 588-590: I would use (i) and (ii) instead of (a) and (b).*

We made the suggested change.

Reviewer #1: *Line 618: Can you add few words to make explicit what specifically you refer to from Zhang et al. (2013)?*

We rephrased to "Similar conclusions were drawn in earlier work by Zhang et al (2013)" to clarify.

Reviewer #1: *Line 653: I guess you meant to refer to Section 5.2? Or have I missed 49 sections? Hope not!*

Whoops, yes this was a latex error that we've now fixed.

Reviewer #1: *Appendix A: Not sure if the journal allows for it but I felt like Appendix A could easily go in 'Supplementary Information', as it is contains mostly technical details not crucial to the paper (in my opinion), but still worth having available if one is interested.*

Yes we actually had imagined that both Appendices would go into Supporting Information in the end. We just left them as Appendices for now to make the material more easily accessible to the reviewers. They have now been moved into Supporting Information.

Comments on figures

Reviewer #1: *Figure 1: Why are you using dashed lines for models? A solid, differently coloured line would look neater I think (similarly in Figures 2, 11, and 14).*

The dashes mean the lines can be unambiguously distinguished irrespective of any colour-blindness or other impairments of sight. Experience suggests this can be helpful for some readers so we'll leave these as is.

Reviewer #1: *Figure 4: I don't think I follow how panel (e) is constructed, it should be showing the areas that have common significant correlation across panels (a,b) here and Figure 3 (a,b), but Figure 3b does not feature such a wide region of significant correlations.*

Sorry, there was a silly mistake in the code generating Figure 3. The stipling in Figures 3(a) and (b) were not showing significance of ERA20C gridpoints alone, but of some intersection of significance across ERA20C and ASF20C. This was due to an earlier version of the figure which hadn't been properly updated to reflect the final choice of presentation. The code generating the intersection of correlations used for panel (e) in Figures 3 and 4 were both using the correct masks, it was just the masks highlighted in Figure 3(a)/(b) that were set incorrectly.

We've fixed this now and the new Figure 3 shows stipling more clearly in line with what you see in Figure 4(e). Thanks for spotting this!

Reviewer #1: *Figure 5: To retain some information on the magnitudes of the SST anomalies I would not normalise it but simply use a second axis on the right and rescale it to match the current positions. I would also include the value of the*

standard deviations of the three jet speed indices in the legend, next to the correlation score (C=…, /sigma = …)

We added a 2nd axis to show the SST magnitudes, but had to keep the jet speed indices normalised, since ERA20C and forecast jet speeds have different magnitudes. We also made a similar change in Figure 1, 2 and S1, to show explicitly the different magnitudes of the jet speed of ERA20C vs forecasts. We also added the sigma values in the legend as suggested.

Reviewer #1: *Figure 8: Can you modify the colour bar to avoid allowing values of C outside the range [-1, +1]? Similarly in Figures 3, 4, and B4.*

We made the suggested change.

Reviewer #1: *Figure 10: You do not really discuss panel (b) in the text, either do that or remove the panel.*

We added a line commenting on Figure 10b (9b in the revised), namely that comparing it to Figure 9d (8d in the revised) confirms that the association between winds and the SPNA projects onto the jet speed signature.

Reviewer #1: *Figure 12: You have plotted a box which seems to correspond with the SPNA but it is not described in the caption.*

We've clarified in the caption that the SPNA is highlighted with a box.

Reviewer #1: *Figure 13: As mentioned in another comment, specify what you mean by positive/negative heatflux.*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #1: *Figure 15: Are the units in panel (a) correct? Also, the ticks look a bit funny, I would use integers only (1, 2, 3,…, 12). As for panels (b) and (c), the red line is the same right? Could you not just make it one panel, given that you normalise the time series anyway?*

The units were indeed misleading: they are actually ppbv (parts per billion by volume), and we've now made that clear, along with editing the ticks.

We have also followed the reviewers suggestion and collapsed the figure into 2 panels.

RESPONSE TO REVIEWER #2

GENERAL COMMENTS FOR BOTH REVIEWERS
We thank the reviewer for their thoughtful comments and feedback. Based on these comments and those of the other reviewer, we have carried out considerable revisions. We provide a summary of the biggest changes:

- Several sections have been shortened or cut entirely (such as the section on stochastic atmospheric forcing) in an effort to make the paper more concise. The number of figures has also been reduced. Note that Section 6 was not cut entirely, as the second reviewer suggested. The analysis of aerosols and the AMOC have been retained in the revised manuscript, since we felt these added value to the paper (see responses to 2nd reviewer).

  We welcome further feedback from both reviewers (and the editor if they are reading) on the value of the trimmed down Section 6.

- The discussion has been largely removed from the various results sections and is now relegated to its own section close to the end.

- The statistical significance testing has been modified somewhat, so that we now consistently adopt a null hypothesis where atmospheric timeseries are modelled as white noise and SSTs using the Fourier phase shuffle method. Details of the tests and thresholds used are now more consistently made clear, including in the captions of figures.

- Analysis of heatflux variability, such as gridpoint correlations between SSTs and heatfluxes, have been removed. Based on comments from Reviewer #2 and discussion with co-authors and other colleagues, we now feel that much more careful analysis would be required to compare the heatflux variability across the different datasets we use.

  The related analysis where we compute correlations between daily SSTs and surface temperatures in the SPNA region has been retained, as we judged this was less prone to the subtleties associated with the heatflux variability. The speculation on the role of roughness lengths has been kept, but moved to its own section in the Discussion and is phrased more cautiously.

We also need to flag an error in the original Table 2 (correlations between SPNA and JetSpeed). The values for ERA20C and ASF20C reported were based on a slightly different SST domain corresponding to the Labrador sea, due to the fact that much earlier versions of this work focused on that region. The corrected correlations using the SPNA domain are now reported: the reviewers will note that they are very similar to the previous values, and that questions of statistical significance are not altered (which is why we had failed to notice this before). Apologies for this!

All the revised parts are highlighted in RED in the revised manuscript.

We now respond to each comment in turn.

Major comments:

Reviewer #2: *Maybe the original contribution is the focus of the jet speed predictability and the analyses of the signal to noise from different datasets. The figure 6 and the related text, the section 4.4 and figure 7, the section 5.1 and figure 8, the section 5.2 and figure 11, or the section 6 seems somehow be out of context and unnecessary. The minor comments below provide more details.*

We agree that Section 6 could potentially be removed. However, on balance, we think it adds some weight to our assertion that the SPNA is a plausible source of forecast skill, because it helps reinforce that there isn't a super obvious alternative source of skill, like sulphate aerosols, and that SST variability alone can account for what we see (as in the AMOC plot). Given that the jet speed perspective is new, and we have argued it could make signals more transparent (compared to the NAO), it also seems of immediate interest to re-examine aerosols and the AMOC in this new context. We have however now removed the subsection about stochastic forcing, which was probably less interesting and relevant.

Section 5.1 and the associated figure could potentially be replaced by something else, but some justification is needed for the computation in Section 5.2, which we think is important to provide evidence for the tropospheric pathway (see our response to minor comment below for more on this computation). Section 5.1 provides that justification nicely using an independent line of inquiry, which thereby helps support the paper overall.

The other sections/figures that the reviewer mentions are ones we do feel are important. Please take a look at the revisions we have made in response to the minor comments and let us know if you still think these are not of interest.

Reviewer #2: *The statistical analyses can be limited to only the most robust test. Simple and not-adapted statistical tests should not be shown and discussed. In the*

manuscript several statistical models are used, many times such models are not adapted. Confidence intervals are mixed up with threshold values. The level of statistical significance is not always specified in the figures panels. The minor comments below provide more details.

The statistical models have been cleaned up, with Fourier phase shuffling used to represent the SPNA in all cases now. References to confidence intervals have been replaced with references to thresholds: apologies for this confusion of terminology. Statistical significance levels are now shown clearly in all relevant panels/captions.

Reviewer #2: *The discussion should be limited to the interpretation of the results shown in the section dedicated to results. The authors can also provide the larger picture and discuss how their results compare with other recent studies. However, I have the feeling that the manuscript contains too much discussion within the results section. The minor comments below provide more details.*

We have now created a separate Discussion section, and moved as much of the discussion that seemed reasonable from the results sections to the discussion section. Please let us know if the balance seems more reasonable now. We appreciate that we have many points we want to discuss.

Reviewer #2: *The heat flux from ERA20C reanalysis is resulting from a model using prescribed SST. This would somehow modify the negative heat flux feedback, as found by Barsugli and Batisti (1997). Therefore, the results from ERA20C have some limitations that could be discussed.*

We thank the reviewer for their points concerning subtleties around the interpretation of heatflux variability in ERA20C. Based on this and further discussions with colleagues and co-authors, we decided we were no longer confident in the interpretation of the heatflux diagnostics we had computed (e.g. correlations between gridpoints and SSTs). The heatflux analysis has therefore been removed, as stated in the Response To Both Reviewers.

Minor comments:

Reviewer #2: *L74: A SST-prescribed simulation is different from a nudged forecast.*

It should have read as follows, and now does (the key change has been underlined: it used to refer to an atmosphere-only hindcast):

“While taking decadal averages of a seasonal hindcast obviously does not constitute an actual decadal *forecast*, it nevertheless turns out to be useful to think of it as

being like a `nudged' forecast, where both the atmospheric and oceanic state are being nudged back towards observations at the start of each winter (and moreover, for ASF20C, the SSTs are being forecasted perfectly)."

That is, the comparison to a nudged forecast is made more broadly in the context of a generic seasonal hindcast (coupled or uncoupled), and if one has additional prescribed the SSTs (like in ASF20C) then it's akin to having perfectly forecast the SSTs.

We hope the edit has made this analogy clearer.

Reviewer #2: *L78-80 : I do not understand. Simplify the causality of what?*

We have edited to say "the lack of coupling in ASF20C simplifies the question of causality between ocean and atmosphere".

Reviewer #2: *L144 : change to ''the CMIP6 historical experiment for 31 CMIP6 model. Only one member was selected for each model."*

We rephrased these lines as follows, in line with the reviewer's suggestion:

"We analyse the CMIP6 historical experiments for 31 CMIP6 models \citep{Eyring2016} detailed in Table A1 of Appendix A: only one member was selected for each model. We also analyse a total of 71 ensemble members from 28 distinct CMIP5 models \citep{Taylor2012} detailed in Table A2."

Reviewer #2: *L147: change to '' with initial conditions sampled from a free-running pre-industrial control run". The pre-industrial simulation is not considered a spin-up. The spin-up determines the initial conditions of the pre-industrial run in the CMIP protocol.*

We made the suggested change.

Reviewer #2: *L151: ''the same historical forcings as CMIP6" I believe the forcing is different from CMIP6, at least for aerosols or land-use. Please check.*

You're right, we had forgotten about these differences. For these coupled simulations, the only difference is in aerosols and land-use, as you note: this can be verified from Table 1 of Haarsma et al (2016), which summarises the differences.

Reviewer #2: *L166 : ''along with any linear trend" I am not sure that one should remove the linear trend. Are the results sensitive to the removal of a trend?*

See the “general comments to both reviewers”.

Reviewer #2: *L172 : ‘’peaks” Do you mean is maximum?*

Yes: we rephrased to reflect this.

Reviewer #2: *L191-193: I disagree with the statement. Significant differences between ASF20C/CSF20C and DPLE could be searched. The difference between the two datasets cannot only be explained by the different sizes of the data.*

We have rephrased this as follows:

“Finally, in cases where ASF20C/CSF20C and DPLE disagree on whether an effect is statistically significant or not, we always choose to place higher credence in ASF20C/CSF20C, due to the fact that (a) their sample size is approximately twice as large ($N=109$ vs $N=56$); and (b) they have more ensemble members (51 vs 40). It is possible that such disagreements reflect genuine differences in physical mechanisms, but we do not consider this possibility for the present.”

In other words, we agree that there might be genuine differences between DPLE and ASF20C/CSF20C (and we mention one such possible difference elsewhere), but we will not consider this in the current paper, for the simple reason that this would further complicate what is already a long paper. We hope the reviewer will accept this limitation and our clarification on the point.

Reviewer #2: *L197-198 : ‘’due to the signal-to-noise paradox” . I do not understand. Can the authors provide the scaling applied for ERA40C and ASF20C. This is not the signal to noise paradox, but reflects the fact that ASF20C is obtained with an average of several members, which should damp the variability.*

In response to reviewer 1 we have added separate axes for reanalysis and models, to demonstrate this effect: this has been done for Figures 1, 2 and B1. This allows for an easy comparison of the relative magnitudes involved.

It is true that in general an ensemble mean will have smaller magnitude than obs, but this effect is greatly pronounced here as a consequence of the paradox. We note that the paradox is, by its definition, a statement related to correlations and not ensemble mean magnitudes. However, the statistical models that have been devised to explain the paradox (such as that of Siegert et al 2016 and Strommen and Palmer 2019) suggest that the ensemble mean magnitude would be much higher if the model bias encoded in the given statistical model were eliminated.

Besides adding the extra axes, we have rephrased to say that the small magnitude is “one aspect” of the paradox.

Reviewer #2: *L201-202: What is a 10/30-year correlation?*

We rephrased to say “The correlations obtained using 10/30-year smoothing” instead.

Reviewer #2: *L202-203: How can the authors speculate the skill of decadal forecast from the results of ASF20C? ASF20C used forced SST. The mechanism leading to the persistence of SST in the early twentieth century could be different. The external forcings are also different.*

We agree that we don’t have enough reason to express this claim with great confidence. We have therefore rephrased to the weaker statement: “suggesting that the decadal forecast skill they reported might extend all the way back to 1900.”

Reviewer #2: *L205-206 : I do not understand the analysis and the conclusions here. NAO and jet speed are related. But is it different for the interannual variability? I guess this is model dependent as well.*

We are trying to support the stated claim that “all of the skill that ASF20C has at reproducing decadal NAO variability can be accounted for by the jet speed”. To do so, we want to remove the component of the NAO variability associated with the jet speed. We do this by regressing jet speed against NAO:

NAO = a*JetSpeed + b + residual

(with constants a and b). The residual is then interpreted as the component of NAO variability independent of the jet speed. We find that ASF20C has little/no skill at predicting this residual. Since ASF20C can predict the NAO, and also the jet speed, we conclude that it is the jet speed that is responsible for the skillful NAO prediction.

This does depend on the assumption of a linear relationship, but this is justified by the fact that the NAO and Jet speed are to very good approximation jointly Gaussian, as noted in previous studies.

Reviewer #2: *L206 : Are the values 0.1 and -0.1 statistically significant? The p-value could be small given the large ensembles used.*

They are not significant. Indeed, the residuals are also normally distributed, so the null hypothesis of the residuals in ERA20C and ASF20C being independent white noise generates essentially identical significance thresholds to those shown in Table 1.

Reviewer #2: *L207: ‘’can be accounted for by the jet speed” Maybe ‘’is related to the jet speed” seems a better formulation.*

We made the suggested reformulation.

Reviewer #2: *L211-212 : ‘’is much less sensitive for DPLE” I do not understand. What sensitivity is discussed here?*

The reviewer appears to have misread the word “sensible” as “sensitive”. We are saying it is not very sensible to take 30-year means for the DPLE, since the DPLE only spans 56 years.

Reviewer #2: *L212-213: ‘’is likely to just be noise” Can the authors test the correlation to see the related p-value. This statement is too qualitative and needs to be verified.*

Yes, this was lazy of us. We have now verified that the correlation is not significant using the same null hypothesis as used in Table 1 (line 231 in revised).

Reviewer #2: *L214-215 : ‘’The lag-1 correlation is approximately zero” What lag is used here? Which variable leads? Do the authors mean the lag-1 autocorrelation? What is the unit for the lag? Lag=1 day? 1 month? 1 year? What do the authors mean by approximately? With such a large sampling, even a small autocorrelation can be significant.*

Yes, we mean the autocorrelation at lag 1. The units are in years, as suggested by our phrasing “interannual lag-1 correlation”. By “approximately zero”, we mean 0 to 1 decimal place. For example, in ERA20C the lag-1 autocorrelation is 0.039, which is not significant.

To make both points clearer, we have rephrased to say “the 1-year autocorrelation is approximately 0.0”.

Reviewer #2: *Table 1 : for conciseness, the authors might provide the correlation obtained and the related p-value. I do not think that the values for the 90% and 95% confidence levels are needed here.*

We appreciate the appeal of consistency, but our experience is that p-values by themselves can be misleading or misunderstood. This viewpoint has been raised by several scientists over the years, including in the atmospheric sciences, such as work by Ted Shepherd: https://link.springer.com/article/10.1007/s10584-021-03226-6

This is why we have tried to be very explicit about the statistical testing, by (a) bootstrapping explicit and simple statistical models where feasible and (b) reporting actual confidence levels/significance thresholds as opposed to just p-values. Our experience is that these are more intuitively understood by the average

climate/atmospheric scientist. We hope that the reviewer will therefore not object to us leaving the table as is.

Reviewer #2: *L221: ‘’within rho < 0.01 and not rho < 0.05” What is rho here? A correlation? A p-value? Perhaps p is a better notation than rho. I am confused as both confidence level (with large value if statistically significant) and significance level (with small value if statistically significant) are discussed. It would be more clear to stick with one convention (either significance or confidence level).*

The rho is meant to signify a p-value. We appreciate this can be confused with notation for correlation, so have changed all occurrences of rho to a p.

Reviewer #2: *L229 :”it is therefore plausible […] shift rho to below 0.05 ‘’ What is a shift in the null hypothesis? This is a speculation unless the authors do report the results of another analysis.*

We have rephrased this as follows:

“Our results here contrast somewhat with those of \citet{smith2019robust} and \citet{athanasiadis2020decadal}, which both report statistically significant decadal NAO forecasts with $p < 0.05$. However, their significance tests differ from ours, and the smoothing they use is also very close, but not identical, to the 10-year running means we use. Table 1 shows that the 10-year ASF20C and DPLE jet speed correlations sit neatly between the bounds of the 5 and 10\% significance thresholds, and it is therefore plausible that these differences can explain why they report a p-value less than 0.05 and we do not.”

It’s true that this is speculation, since we did not repeat our tests using the exact methodology of the two cited papers. However, it seems reasonable and we also are clear to mark it as speculation (“it is plausible that ….”) and not fact.

Reviewer #2: *L231-235 : I do not believe this discussion is needed at this point. I suggest that the authors remain on the description of their results.*

We have moved the first line (“For this article, we will assume that ASF20C and DPLE really do have significant skill at predicting decadal variations in the jet speed”) to the end of the previous paragraph, because it summarises the discussion about the significance of the skill.

The rest of the lines have been removed, as suggested by the reviewer.

Reviewer #2: *L233 : ‘’significance” -> “confidence level”*

The relevant line was removed, as per the previous point.

Reviewer #2: *L237-255 : This discussion is confusing. I am not sure it is needed. I suggest that the authors stick with a description of the results and move the discussion in the dedicated part of the paper.*

We respectfully disagree with the reviewer here, and think some motivation of the sort given in this subsection is necessary. In the analysis that follows, we are making crucial use of the assumption that (a) the signals we are looking for have to be visible already on seasonal timescales and (b) the mechanisms involved do not involve within-season coupling.

For example, point (a) is the key assumption that allows for the creation of Figures 3 and 4, and these figures are what we use to justify the assertion that the SPNA is the only plausible source of an SST-based signal. Point (b) is a key point used to justify arguments concerning causality between the ocean and the atmosphere, since ASF20C is uncoupled.

All these arguments and analysis methods would be unjustified without first making these assumptions, so we believe it is important to state these up front, as this subsection aims to do. We have kept this subsection for this reason, but have tried to simplify the discussion to focus on the key points. We hope the reviewer will find it more readable now.

Reviewer #2: *L237: ‘’The similar behaviour of ASF20C and DPLE suggests” I would not argue that the mechanism is the same because the correlation is positive with ERA20C in both cases.*

It’s not just that the correlations are positive in both, but (1) that they’re similar in magnitude and (b) that in both forecasts it’s only the jet speed which is predictable. It certainly can’t be excluded that they attain this similar behaviour through different mechanisms, but our interpretation seems pragmatic and reasonable. The possibility that they really might have different mechanisms is a caveat that we already raised explicitly in the conclusions.

Reviewer #2: *L249: ‘’any given initialized ASF20C atmospheric forecast has no direct knowledge of this” I do not understand the discussion here. Do the authors argue the forecast are not influenced by the decadal variability of the prescribed SST?*

Yes: the atmospheric variability simulated in ASF20C is purely a function of forcing taking place within a single season (since the ASF20C forecasts only span a single season). We have edited this subsection already based on earlier comments, and hopefully the points made are clearer now.

Reviewer #2: *L254 : ‘’deficient coupling may play a role” This seems to be a speculation or needs to be discussed with appropriate references.*

We added a reference to a paper where this is explicitly conjectured.

Reviewer #2: *L263 : ‘’stratospheric variability in ERA20 is highly biased” :Can the authors be more specific? What are the biases in variability? Do you the authors mean that the observed stratospheric variability is poorly reproduced? We cannot exclude some impacts of the stratosphere when assimilating only surface variables.*

We have added details about what is meant, as also requested by Reviewer 1. It now reads as follows:

“It is possible in principle that the skill comes from the stratosphere \citep{omrani2014stratosphere}, which has a decorrelation timescale of several months. However, the initial conditions of ASF20C (CSF20C) come from ERA20C (CERA20C), which only assimilates surface variables. Because these do not strongly constrain the stratosphere, the stratospheric variability in ERA20C is highly biased: \citet{o2019importance} found that the amplitude of the quasi-biennial oscillation (QBO) is substantially weaker in ERA20C compared to reanalysis products that assimilate the atmosphere. Furthermore, they found a greatly reduced downward propagation, and the seasonal timescale association between the QBO and the NAO in hindcasts initialised with ERA20C was essentially zero. We interpret this as evidence that there is limited scope for skill from stratospheric initial conditions, and therefore choose to not consider potential stratospheric sources of skill in our analysis.”

We also pointed out that Simpson et al. (2018) included an independent argument suggesting that the stratosphere is not playing an important role in decadal jet variability.

Reviewer #2: *L265 : ‘’has been examined previously (O’Reilly et al., 2019)” Can the author summarize the results? Can the stratosphere explain the skill associated with the decadal time scale?*

See above.

Reviewer #2: *L270 : ‘’exclude ice from our analysis ‘’ The authors should note that many studies have noted that sea ice has only weak impacts (Liang et al. 2021; Ogawa et al. 2018). The authors can note that sea ice was not observed before 1979, as the impact of sea ice is therefore not known for 1900-1979.*

We added the references and commented on sparsity of observations prior to 1979.

Reviewer #2: *L272 : ‘’or locally” Note that the role of aerosols has large uncertainties. I do not believe that one can exclude a global signature for their forcing.*

This has now been subsumed into the new subsection 7.1 of the Discussion, where this is discussed more fully.

Reviewer #2: *L258-277 : This discussion can be shortened. First the authors can discuss why they focus on the SST based on previous works. Then the authors can list the other possible drivers. But most of the discussion on the potential role of other drivers, should be moved to the discussion-conclusion section.*

We did as suggested: see the new Section 7, and in particular 7.1.

Reviewer #2: *Figure 3 : as argued by the authors, the student t-test is not robust. Such a test cannot be shown. I do not see the interest of a ‘’first pass” test. For conciseness, I recommend only showing the results of a robust test.*

The test used is deliberately weak, which we have now explained in the revised paper. A very stringent filter for common signals, e.g. by demanding significance (p=5%) with a Fourier phase shuffle test across all timescales and datasets, has the benefit that anything which survives the filter is definitely a very robust signal. However, the downside is that it is arguably *too* stringent. Statistical significance thresholds are inherently arbitrary, and a region which passes a 10% threshold but not a 5% threshold may still be a source of predictability in the forecasts: similarly with regions which pass the 5% threshold for a weak test but not a strong test. In other words, stringent filters may be criticised as having high false negative rates.

Regions which survive a weak filter (like the one we impose) are, on the other hand, not clearly robust, but the flip side is that it is easier to argue that regions which *don’t* survive the filter are unlikely to be important (low false negative rates). What we find here is that even when applying this weak filter (due to the weakness of the t-test in the presence of large SST autocorrelations) we still find that the SPNA is the only obvious region which survives. We think this constitutes strong evidence that other regions are not important. Strengthening the filter considerably would make this less clear.

We have therefore left the filter as is. The above reasoning is now included in the revised paper (paragraph starting L277 in revised).

Reviewer #2: *Figure 3 : The authors show a plot with grid point with significant correlation in ERA20C and ASF20C. At each grid point, four tests are applied. The statistical significance is then different from panels (a), (b) or (c). The Holm’s correction could be applied when conducting multiple tests.*

We agree this would generally be a good idea, but we think in our case it is not necessary. Our procedure is not meant to determine true positives, only to get rid of true negatives. We use a weak test deliberately for this reason (see above reply). Then we apply a proper, stringent significance test to the regions that survive the filter (which turns out to only be the SPNA).

Holm-Bonferroni would make the filter stronger, and weed out false positives. This would be crucial if we were to conclude from Figs 3 and 4 alone that the SPNA has a robust link with the jet speed. But since we perform a stringent, independent test of this link after Figures 3 and 4, this isn't crucial in our case. We therefore have decided to not add a Holm-Bonferroni correction.

Reviewer #2: *Figure 3: please mention the level of statistical significance in the legend.*

Done.

Reviewer #2: *L295-304 : The authors should describe the important differences between Figs. 3bd and Figs. 4 bd, or between Figs. 3ac and Figs. 4ac. In all cases, I agree that there are cold anomalies in the SPNA, but there are associated with different SST anomalies elsewhere. For instance, DPLE shows strong negative SST anomalies in the equatorial Pacific for 1yr correlation. 10-yr correlations are associated with near-global cooling in the case of CSF20C, and near-global warming in DPLE or ERA20.*

While we agree that the differences are interesting, they do not seem to be clearly relevant to the study at hand. Given the reviewer's major comment concerning the length of the manuscript, we are hesitant to add further discussion here. We have therefore rather added a line at the end of the discussion on Figures 3 and 4 stating that while the differences are interesting, they are not clearly relevant and hence not discussed.

Reviewer #2: *L302-303 : ''we consider it justified to weight these datasets higher than DPLE". I disagree, if the statistical significance is correctly calculated, I do not see why the results of DPLE should be excluded or weighted.*

p-values are always influenced by the sample size: longer samples reduce the impact of noise and hence tend to produce smaller p-values. If a shorter sample gives non-significance but a longer sample of the same data gives significance, it is therefore generally reasonable to give higher credence to the longer sample, and this is clearly common practice in the field.

There is no clear way to adjust the p-value for sample size without making prior assumptions about the processes involved, which we want to avoid doing at this point since we are testing across thousands of gridpoints covering the entire range of ocean regimes.

We have therefore left this as is.

Reviewer #2: *L305 : ‘’the SPNA is the only plausible […] region” : I would be more careful and argue that it is possibly an important region.*

We have rephrased to the following, more cautious statement:
“To summarise the above discussion, the SPNA emerges as a common region of interannual-to-decadal correlations between SSTs and the jet speed across ERA20C, ASF20C, CSF20C and DPLE, and appears to be the only such common region. We interpret this as strong evidence for the importance of the SPNA in driving decadal jet predictability.”

Reviewer #2: *L312-317 : I do not understand this discussion. What is the point of the discussion?*

We have searched for significant correlations across a huge number of predictors (every gridpoint). This is a situation where the false discovery rate is expected to be high: if one tests enough predictors some are bound to be significant by chance. We are simply trying to argue that this doesn’t constitute a strong argument for rejecting the significance of the SPNA gridpoints. We have in any case somewhat rephrased this discussion in response to Reviewer 1, and we hope you will find it more clear as a result.

Reviewer #2: *Figure 5 : ‘’(i.e. SSTs are drawn […] in the case of DPLE jet speeds)” this can be suppressed for conciseness.*

We made the suggested change.

Reviewer #2: *Figure 5 : can the authors provide the scaling applied to each if the time series? Can the SST from DPLE be added to the plot?*

As with Figures 1 and 2, we have added a 2nd y-axis to show the magnitude of the SSTs and jet speed separately: the jet speed timeseries have all been normalised due to the different magnitudes of reanalysis and forecasts. We are hesitant to add the DPLE SSTs to the plot, given that it is already crowded.

Reviewer #2: *Figure 5: the time series DPLE speed seems to have a variance different from one.*

We double checked that it does, in fact, have a variance of 1.

Reviewer #2: *Table 2: the intervals given are not confidence intervals, as such intervals are centered on the actual correlation. The intervals given are thresholds corresponding to the significance level of 5%.*

You're right, thanks for the correction. We've reworded to say 5% thresholds, both here and elsewhere.

Reviewer #2: *In the case of Table 2, only the test from Ebisuzaki (1997) is relevant. The SPNA SST shown in Fig. 5 cannot be considered as an AR1. The p=5% thresholds are also not needed if significant correlations are already shown in bold.*

L322: ''with a AR1 process " The SPNA SST shows a pronounced multi-decadal variability. An AR1 seems not appropriate.

Yes, fair enough. We have removed the Simple Model test from this section and Table 2. We now use the following null hypothesis here. The jet speed is modelled as a normal distribution with no interannual memory (as before). The SPNA is modelled with its full autocorrelation structure, by using the Fourier phase shuffle method. By drawing 10,000 random such timeseries for each, and correlating them and their smoothed versions, we estimate 5% thresholds for each timescale.

The thresholds thus obtained differ somewhat from those originally included, due to the fact that in the original version the jet speed was also modelled using the Fourier phase shuffle method. However, since the jet speed has ~0.0 interannual autocorrelation, there is no clear physical basis for baking in the higher lags into the null hypothesis. This new null hypothesis is also more consistent with the significance test used to assess decadal skill.

Either way the main conclusions don't change. ERA20C is significant everywhere, while the forecast models are significant at some timescales and not others.

Reviewer #2: *L332 : ''while considerable" What do the authors consider such values as large? Maybe the authors should only discuss the statistical significance here.*

We removed the phrase "while considerable".

Reviewer #2: *L335 : ''this amount of smoothing is unlikely to be sensitive here" This sentence and the related footnote is speculation. I believe this is not needed.*

We removed it.

Reviewer #2: *L335-340 : I believe that the discussion is not needed here.*

We have cut the discussion down to just making the point that significance needs interpretation and physical reasoning needs to be included in any final judgement: we cite Shepherd 2021 here.

Reviewer #2: *L354 : ‘’a weaker jet should transfer less heat into the ocean” I suggest changing the formulation. The climatological heat flux is out of the ocean in the SPNA, i.e. it is upward. A weaker jet decreases such upward heat flux, with anomalous downward heat flux.*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #2: *Please add the sign convention from the heat flux in Figs. 12 and 13.*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #2: *L359 : ‘’ different instances of stochastic atmospheric variability ‘’ The authors should note that the initial conditions and all boundary conditions could also be included in the s term. The residue is here attributed to the stochastic forcing.*

We rephrased to “with s representing a forced signal from the initial conditions and boundary forcings (e.g. SSTs)”

Reviewer #2: *L363-364 : The authors could note that this dipole is well known to be the NAO SST signature.*

We now do.

Reviewer #2: *L364-365 : ‘’with correlations of around -0.06 (note the color scale)” The 5% significance should be shown and only discuss the significant correlation. In Fig. 6, the correlations are locally significant.*

We added stipling for significance to the figure and note that the SPNA regions have almost no significant gridpoints. Note that this analysis is now in Section 7.2 of the discussion (paragraph starting L586).

Reviewer #2: *L356-366 : I do not understand the purpose of this analysis. It is well known that the tropospheric memory is small (10 or 15 days), so any significant ensemble mean anomalies should represent the effect of boundary conditions. The forcing from the stochastic NAO is maybe out of the scope of this study. It can also be estimated using lag correlation or regression.*

There is an ongoing debate in the literature concerning the role of ocean-atmosphere coupling in the North Atlantic. In particular, are phenomena like the AMV just a result of stochastic atmospheric forcing being integrated by the ocean, or is the AMV also driven by ocean dynamics? The former viewpoint takes the strong view that decadal timescale ocean variability is causally driven by the atmosphere. Here, we have reported correlations between the atmosphere and the ocean, and so proponents of such a viewpoint may well argue that these correlations are just showing the causal influence the atmosphere has on the ocean. Our analysis is meant to provide a quantitative reason to believe that the correlations we show also include the causal influence of the ocean on the atmosphere.

Note that we agree with the reviewer that the best argument here is the apparent existence of predictability, which as the reviewer notes, essentially *must* come from something like the SSTs.

We do agree that a more proper estimate of the stochastic forcing is outside the scope of the study, this objection is quite common in our experience, so it seemed worth including this analysis to show explicitly that it seems hard to explain what we find if one assumes there is no causal influence from the ocean.

Reviewer #2: *L362 : ‘’the noise terms” change to ‘’the terms unrelated to initial or boundary conditions”.*

We rephrased to “ the link between the SSTs and the atmospheric variability unrelated to initial or boundary conditions”

Reviewer #2: *L369 : ‘’is weaker than” Note that the correlation cannot be used to quantify the magnitude of the signal. Maybe the authors can calculate the regressions, which can quantify the change in °C per m s-1.*

We have rephrased to say “the link between the SSTs and the stochastic forcing is notably weaker than for the ensemble mean”, to emphasise that the correlations are measuring the ‘link’, not the forcing per se.

Reviewer #2: *L366-369 : I do not see the point of adding this new analysis.*

As noted in a previous response, the idea that decadal SST variability is actually a result purely of stochastic atmospheric variability is not an uncommon position in the literature, and such a position would challenge our interpretation of decadal timescale SST-Jet correlations as demonstrating a causal influence of the ocean. It therefore seemed worth tackling this potential objection head on.

Reviewer #2: *L375-383 : Again, I do not see why the SPNA SST would be an AR1 given the large multidecadal variability in this region.*

The SPNA is now modelled using the Fourier phase shuffle method.

Reviewer #2: *L372-399 : I do not understand what is investigated here. The authors present that they would like to ‘’test whether the forcing taking place on seasonal timescales suffices to explain the bulk of the forcing taking place on decadal time scale”. What do the authors mean by forcing? I believe that the time series in Fig. 5 does not include any seasonal variability… most of the variability is at the decadal time scales. In Figure 7, surrogate time series are generated using an AR(1) for the SPNA SST. What is the null hypothesis then?*

We are proposing that the decadal predictability of the jet speed is due to forcing from the SPNA to the jet taking place on seasonal timescales. The evidence for this proposal has so far relied on the assumption that ASF20C is representative of what's happening in the real world and in DPLE (a genuine decadal forecast). The analysis here is intended to provide additional support for our proposal, by showing that the correlations seen on multidecadal timescales are consistent with what is expected from integrating the seasonal timescale link over time.

We have now edited the analysis to use the Fourier phase shuffle method for the SPNA, instead of the AR1 assumption.

The null hypothesis is as follows. The interannual SPNA is modelled with its full autocorrelation structure, by using the Fourier phase shuffle method. The interannual jet speed and SPNA are assumed to be linearly related according to a least-squares regression fit. Given this null hypothesis, one can assess what 10 and 30-year correlations can be easily attained. What we show is that the observed 10 and 30-year correlations are indeed easily attainable using this null hypothesis.

We have rephrased the section here. Please let us know if it's still unclear.

Reviewer #2: *L391-396: I do not understand why the authors speculate that atmosphere-ocean coupling would lead to a correlation outside of the correlation obtained from the surrogate time series. Can it mean that the SPNA SST is not an AR1? The uncertainties in a and b need to be accounted for when using a regression from prediction as here. What has been done here?*

Yes, one possibility we were alluding to is that maybe the SPNA is not an AR1, as had been assumed. This is no longer relevant given that the change in the null hypothesis (see above reply) now brings ERA20C within the threshold for all timescales.

About uncertainties in a and b, while these need to be taken into account for a proper analysis of the full range of correlations attainable, all we are trying to do is show that the 10 and 30-year correlations are consistent with the null hypothesis. Adding uncertainty around a and b will only have the effect of widening the distribution of correlations further. However, even without this extra uncertainty, the correlations already fall within the 95% confidence interval. Therefore they will

certainly do so also with an even wider distribution. We thus omit this additional complication in favour of brevity.

Reviewer #2: *Figure 7 : What is the unit in the y-axis? It seems different from the relative frequency (some values >1) or absolute frequency (larger values are expected from a size of 10000).*

The histograms have been normalised so that the area underneath is 1. We now state this in the caption to the figure.

Reviewer #2: *Section 4.5. This discussion should be moved at the end of the manuscript.*

We have done so.

Reviewer #2: *L429: I do not understand the 1980-2015 period. A period comparable to that observed would be better, especially with the importance of decadal/multidecadal variability.*

We aren't looking here to explain multidecadal variability, but rather intermodel spread in climate model means. Using a different, shorter period, but examining a wide ensemble spread, provides a genuinely separate line of evidence towards the importance of the SPNA region, as explained: at 500hPa, the intermodel spread is strongly influenced by temperatures above the SPNA. This is used soon afterwards to perform the geostrophic balance computation.

Reviewer #2: *L430: I am not sure that removal of the linear trend help. Probably the trends are weak and mostly represent internal variability.*

See the "general comments to both reviewers".

Reviewer #2: *The presentation of [Section 5.1] is confusing. Can the authors explain better what was done. From Fig. 8, it seems that the model spread is explored. Did the authors calculate the mean jet speed for each model in 1980-2015 and then calculate a grid point correlation with the corresponding time mean SST in each model? Did the authors calculate the time correlation between the jet speed and the SST? The two are different. Of course, the location of the jet speed corresponds to the different SST biases in models… but this is not connected to what was shown in previous sections. The results linking the SST biases and the jet seems out of the scope here.*

We have rephrased and simplified the section, including restricting the figure to only show the region of the North Atlantic jet. This means we can remove any mention of the Arctic and avoid distractions about other ocean basins.

We agree that it is obvious that the SSTs and the atmospheric temperatures are related to the jet speed in the CMIP models. What is not obvious is which SSTs and which air temperatures. Our analysis here shows that the regions of importance are the jet core and the region around the SPNA. The not-at-all obvious point here is that the tropical Atlantic region (SSTs or air temperatures aloft) doesn't seem to matter that much. This is used to motivate the calculation using geostrophic balance in the following section, and adds an independent line of evidence towards the potential importance of the SPNA.

Reviewer #2: *L436-437: What is removed from the surface?*

We rephrased to refer to the mid-troposphere instead.

Reviewer #2: *L440-441: Can the authors explain what are these planetary wave? How are they connected to the Arctic and the jet speed? Please add references.*

This has been removed, as mentioned.

Reviewer #2: *L442-444 : the correlations obtained are not local. They are wide spread across the Northern Hemisphere. The conclusion and the comparison with the Southern Hemisphere need to be further compared… but again all these questions are not in the scope of the paper.*

This has been removed, as mentioned.

Reviewer #2: *L468-484: I do not see what is the purpose of the analysis here. Indeed, the geostrophy and the thermal wind balance will apply at such large scale. Did the authors expect a different result?*

Geostrophic balance will apply almost tautologically, we agree. However, the calculation we perform is not exactly geostrophic balance, since we have made 3 idealised assumptions:

(1) that the only gradient which matters is the one between the jet core and the SPNA (as opposed to the local gradients everywhere in the jet region);

(2) that the only variations in time that matter are those over the SPNA (as opposed to time variations also in the jet core); and

(3) that the tropospheric air temperature variations over the SPNA are entirely determined by the linear association with the SSTs.

These 3 assumptions are obviously not part of standard geostrophic or thermal wind balance, and it is therefore not obvious a priori that our calculation should agree as well as it does. The fact that it *does* provides some evidence that our idealised assumptions, and hence the tropospheric pathway, is plausible.

We have added a line pointing out that the assumptions made differ from standard theory and therefore there is no reason to expect such good agreement a priori.

Reviewer #2: *L485-494 : this discussion should be shortened and move to the end of the manuscript.*

We have done so.

Reviewer #2: *L495-512 : the discussion can be shortened and move to the end of the manuscript.*

We have done so.

Reviewer #2: *L522: the heat flux at the air-sea interface may also include radiative fluxes. Is it the case here?*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #2: *L523-524 : ‘’correlations between […] heatfluxes and SSTs at every gridpoint” This does not correspond to legend of Fig. 12 when the correlation between the SPNA SST and the heatflux is calculated. Please correct.*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #2: *Figure 12: please indicate the sign convention for the heat flux.*

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #2: *L529 : the authors can note that the results in Gulev et al. show the correlation of the heat flux from observations (another data) and SST at each grid*

point, when using different time filtering. This seems quite different from what is shown in Fig. 12.

This is no longer relevant due to the removal of heatflux analysis.

Reviewer #2: *L532-537 : the fact that the coupling decrease the negative heat flux feedback is well known and has been originally demonstrated in Barsugli and Batisti (1998). This is expected from a prescribed SST in ERA20C.*

Reviewer #2: *L539: Why would the heatflux variability be completely different in ERA20C and ASF20C? Both data set are resulting from prescribed SST. I would show ASF20C heat flux, which is relevant here.*

This is no longer relevant due to the removal of heatflux analysis.

However, we note that the different variability is something we have observed by inspection. For example, the daily seasonal cycle (within the DJF season) is wildly different in ERA20C compared to ASF20C. The latter looks more like what you might expect from the SSTs being an infinite source of heat, while the former looks more subtle and comparable to a coupled model. As far as we understand, the reason for this is that ERA20C is made up of consecutive 1-day forecasts. While each such forecast is uncoupled, the short window means the SSTs don't have time to respond much. The net effect is to make ERA20C look similar to CSF20C and dissimilar to ASF20C. The constraints by data assimilation are probably also having an effect.

These complications are now stated explicitly in the beginning of Section 5.3, to explain why we did not include heatflux analysis and rather focus on just the SST-T2M link.

Reviewer #2: *Figure 13: I do not see why the authors show this figure.*

It has been removed.

Reviewer #2: *L564-595: please reduce the discussion.*

We have.

Reviewer #2: *Section 6 : The drivers of the decadal SST is investigated, but I suggest to remove this part. Why is only the role of sulfate aerosol investigated here? What are the uncertainties in the SO4 emission from ERA20C? Can other aerosols also explain the changes observed? Why do the authors investigate only one CMIP model here? Parson et al. (2020) suggested that the variability in some of the CMIP6 models (such as EC-Earth3) might be too large. Section 6.3 seems to be a discussion on previous results that could be reduced.*

We agree that Section 6 could potentially be removed. We will see if the 1st reviewer has an opinion as well.

Reviewer #2: *Appendix : Can the number of figures in the appendix be reduced? I have the feeling that some of the analyses could be summarized in the text without showing any figures.*

The Appendices were always meant to become Supporting Info before publication. They had been left as appendices to make life slightly easier for reviewers. They have now been moved to Supporting Information.