Response to Review 1

We thank Reviewer 1 for the time taken to thoroughly review this manuscript, which will allow us to improve it further. We particularly appreciate the suggestions to improve clarity on methodology and interpretation of results which the manuscript will greatly benefit from. Our response to Reviewer 1's comments, including specific changes to the manuscript that aim to address them, are in blue below.

Pg2, lines 52-53: [...] To improve clarity for readers here, it would be useful to define the time period of the observational data (presumably 1960/1980 onwards) as well as the timescale of the AMV (decadal).

We have added the word 'multidecadal' before 'timescale of AMV' in the relevant sentence. We have also added '(the last ~45-70 years)' after time period. Additionally, the specific timescales of the observations and reanalyses used can be found in the data section.

Pg3, lines 70-71: [...] It might be useful to highlight why the low resolution version was used here, presumably due to ensemble size limitation for the high resolution model run (and computational limitations for the 20-year runs). In the discussion and conclusions section it might be useful to highlight the evaluation of predictability in a higher resolution model as a potential next step and perhaps speculate on the potential impacts this may have.

We agree that this is important to note. In the same paragraph we have added:

Whilst a higher resolution configuration of the MPI-ESM decadal prediction system exists (Müller et al., 2018), the lower resolution used here facilitated the production of both the increased ensemble size and the extended length runs.

In the discussion, after "*The role of positive ocean-atmosphere and eddy feedbacks for the observed summer atmospheric response to AMV warrants further study.*", we have added the following:

In particular, MPI-ESM-LR has an atmospheric resolution which is coarse relative to other Decadal Climate Prediction Project models (DCPP; Boer et al., 2016), and recent studies have identified that increased atmospheric and/or oceanic resolution can improve the strength of predictable signals due to improved representation of mesoscale eddies (Krüger et al., 2026; Scaife et al., 2019; Wills et al., 2024; Yeager et al., 2023, Zhang et al., 2021). It should be noted that the largest improvements in these studies involve resolutions beyond the range of DCPP simulations; nonetheless whether other decadal prediction systems exhibits yield similar results, and whether there is any resolution dependence on model performance warrants further study.

Pg 4, lines 88-89: “An AMV index is defined as the area-weighted average SST anomaly between 0–60◦N and 280–360◦E”

- If this AMV definition has been used in previous studies, it would be useful to reference these here.

In response to this comment, we have added the following to the relevant sentence:

this index and similar box-average indices are commonly used to define AMV (Deser and Phillips, 2021; Enfield et al., 2001).

Pg 4, line 102: “This method is used here, with the ensemble mean GMSST timeseries extended beyond the end of the historical period”

- It would be useful to define the historical period (presumably 1850/1960-2014) here for clarity.

We have added ‘(1850–2014)’ after ‘historical period’.

Pg 4, lines 119-20: “For this reason, we define East Atlantic (EA) MSLP as the area-weighted mean MSLP anomaly between 45–60◦N and 330–350◦E.”

- Once again, if this is an EA definition which has been used previously, it would be useful to include the reference here.

This definition was chosen as it is where the observed response to AMV is strongest. Previous studies on the (winter or summer) East Atlantic Pattern have used EOF analysis to define it. We have chosen not to as:

- It is a more complex method when the simple box average is suitable for this analysis
- It is dependent on the domain used, the timescale of interest and the dataset (e.g. subjective whether model EOFs should be compared to observed EOFs, or model projections on to observed EOFs to observed EOFs)

To make this clearer, we have added the following **(bold)**:

This region is shown in Figure 2 (c) ***and has been chosen to maximize the observed (HadSLP and ERA5) decadal response to AMV.*** *The same region is also used to define East Atlantic 200 hPa geopotential height anomalies.* ***The results were found to have minimal sensitivity to the box definition (not shown).***

Pg 5, Figure 1 - I wonder whether the positive and negative anomalies would be more clearly visible if the colour bar was set to white (and land set to grey, or similar) for values between -0.05 K and +0.05 K (or -0.1 K and +0.1 K).

We agree that this makes sense and we have implemented this change.

In addition, it would be useful, either in figure titles or the figure caption, to include the periods used for the different observed datasets and models. It would also be useful to define the averaging window and season (presumably JJA means) for the observed datasets and models. Is only the hindcast AMV calculated for the 7-year JJA averaging window? Or are the other datasets' AMV composites calculated in this way?

Thanks for this comment. This was an oversight; we meant to say that for all datasets/variables, 7 year means were taken unless otherwise stated. We have added the following in the Data and Methods section:

For all variables and datasets, 7 year means are taken unless otherwise stated.

Additionally, for all relevant figures, we have now explicitly mentioned this.

In the methods section, we state that all fields are JJA means and we believe that the summer-specific context of the paper is sufficiently clear.

Pg 5, lines 158-159: "the tropical signal vanishes but the extratropical signal remains: a cyclonic anomaly with a position consistent with observations"

- In the different observed datasets, the cyclonic anomaly over the East Atlantic is accompanied by anticyclonic anomalies over Greenland. This is not captured by either the historical or hindcast simulations. Would we expect the models to capture this in any way? Is this an important dynamical feature associated with diabatic heating in the SPG region? If so, it would be useful to mention and discuss in the text.

Thanks for this comment. This is mentioned in the context of ERA5 in the previous paragraph, as the anticyclonic anomaly is comparable to the cyclonic anomaly in that case. We have added:

An anticyclonic anomaly to the north of the cyclonic anomaly is present in HadSLP and ERA5, which is not captured by historical and hindcast simulations.

After the relevant sentence. We do not investigate this further; one simple (but perhaps not complete) hypothesis is that the strong negative pressure anomalies in observations/reanalyses must be balanced by positive pressure anomalies (due to conservation of mass), but this is not evident in the model due to the much weaker pressure anomalies. We are happy to mention this if preferred.

Pg 6, line 161: "To assess predictability, correlation skill for hindcast MSLP relative to ERA5 (Figure 3 (a))…"

- It would be useful, either here or in the methodology, to define the method of correlation used. Presumably this refers to the Anomaly Correlation Coefficient via Pearson's correlation, but this would be useful to define

In this sentence, we have added the following after 'correlation skill':

(defined as the Pearson correlation coefficient between the ensemble mean and verifying observations/reanalysis)

Pg 7, Figure 2

- Once again, it would be useful in the subplot titles or figure caption to define the period (e.g., 1960-2014) used for calculating the MSLP response. Additionally, in the figure caption where: "Composite difference in MSLP during positive and negative AMV using MSLP", this might be better phrased as "Composite differences in MSLP between positive and negative phases of the AMV index". Once again, it would be useful to additionally define the averaging window and season (presumably JJA means) for the observed datasets and models, particularly defining whether there are differences between how the composites are created for the observation-based products (e.g., JJA annual means) compared to the hindcast products (e.g., 7-year JJA rolling means).

Thanks for this comment. There are differences in the time periods: this is how a) and b) are differentiated with the aim of showing how the full and hindcast period differ; ERA5 extends closer to present day than HadSLP; the entire period is used for the historical simulations (cf. response to Reviewer 2's comment on Figure 1 d,e) rather than the period that overlaps with the hindcasts. It would be unwieldy to include all this information in the figure caption and the periods are defined comprehensively in the Data and Methods section.

Pg 7 line 164: "The highest skill levels are to be found off the west coast of Europe, where the negative MSLP response to AMV is found."

- There is a small extension of the skill to the south of Greenland in the HadSLP subplot in Figure 3b. While this is not statistically significant, is it important for predictability? Or is it less relevant as we mostly care about the MSLP response downstream of the diabatic heating in the SPG.

Thanks for pointing this out. We will not mention this in the text as we do not consider it to be too important, although of course this result is evident in the figure and so it is still evident to a reader who might have an interest in this region!

Pg 8, Figure 3

- Once again, it would be useful to include the years over which the skill is computed here (e.g., 1960-2014) either in the subplot title(s) or figure caption.

Thanks for this comment; we have now added this.

Pg 9, Figure 4

- Once again, it would be useful to include the period over which these correlations/regression coefficients are calculated here.

Thanks for this comment; we have now added this.

Pg 10, Figure 5

- Once again. It would be useful to include the period over which these composites are quantified here. In addition, there is a typo in the figure caption, where: “Composite difference in MSLP during positive and negative AMV using MSLP and AMV index from”. Presumably this should be: “Composite difference in 200 hPa geopotential height response between positive and negative AMV phases using the index from…”, or similar.

Thanks for pointing out this error which we have now corrected. With regard to the period, we believe that this information is clear enough in the Data and Methods section and does not need to be repeated here (cf. response to comment on Figure 2)

Pg 10, line 196: “…hindcast response is weaker than in historical simulations, which is consistent with the weaker tropical SST signal”

- The hindcast SST anomalies in the tropical region appear similar in magnitude to the historical SST anomalies in Figure 1e/d. However, when using the ERA5 AMV to define the composites in Figure 1f, the tropical signal is weakened. Could you explain the origin of the “weaker tropical SST signal here”? Is this visible in Figure 1?

and

Page 15, lines 266-269: “However, tropical SST anomalies associated with AMV are weaker and hence less accurate in the hindcasts compared to the historical simulations, and whilst the SPG response in hindcasts is present when defining composites using ERA5 AMV phases (indicating predictability), the tropical SST response largely vanishes.”

- Relating to my previous comment on this here, in Figure 1d/e I do not see how the tropical SST anomalies are weaker for the hindcast in Figure 1e than for the historical simulations in Figure 1d. It would be useful to have some clarification here.

We agree that there isn’t sufficient evidence to claim that the tropical SST anomalies are weaker in the hindcast compared to the historical simulations, and so we have removed the clause *“which is consistent with the weaker tropical SST signal”* from the relevant sentence. We have changed the second mentioned sentence to:

Whilst the SPG response in hindcasts is present when defining composites using ERA5 AMV phases, the tropical SST response largely vanishes, indicating predictability in the extratropics but not in the tropics.

We do not believe that this significantly alters the overall conclusions.

Page 11, Figure 6

- Again, it would be useful to define the fixed time period over which the standard deviation values are computed here. Where the dashed green line is described in the figure captions here: “Solid green lines show correlations between tropical and extratropical AMV against rolling window length, while the dashed green line in (d) shows the same for the hindcast ensemble mean.”, does this mean that the solid green lines in c) and d) show the correlation between tropical AMV and extratropical AMV in ensemble members? If so, how is this computed?

We have added a sentence to the figure caption to explain the time periods used, with reference to the Data and Methods section.

We have changed the relevant sentence to say:

Solid green lines show correlations between tropical and extratropical AMV against rolling window length (calculated as the mean of correlations from individual members), while the dashed green line in (d) shows the same for the hindcast ensemble mean.

Page 12, lines 220-221: “In all cases, the correlation tends to increase with increasing rolling window length.”

- Why does the correlation between the tropical and extratropical components increase with rolling window length? Is this a response that we would expect to see?

We have modified this sentence to explain the likely cause of this, with references:

In all cases, the correlation tends to increase with increasing rolling window length; this may be explained by direct forcing of the tropical Atlantic by the extratropical Atlantic on multidecadal timescales (Brown et al., 2016; Drews and Greatbatch, 2017; Senapati et al., 2024), with remote forcing driving interannual variability (Cai et al., 2019 and references therein).

Page 12, lines 226-227: “The correlation curve for the hindcasts appears ‘noisy’ despite the large ensemble size; this is explained by a significant contribution from the shared ensemble mean component.”

- Firstly, which correlation curve is being referred to here? The solid green line (for member relationship, see above), or the dashed green line (for the ensemble mean relationship)? Secondly, I am not clear on the meaning of ‘shared ensemble mean component’ here. Does this refer to the fact that this ensemble mean contains both tropical and extratropical AMV signals? It would be useful to expand upon and clarify this further.

What was meant that the solid green line appears noisy, as although it is composed of 80 members, those members are not independent of each other, and the more that the

member variance is explained by the ensemble mean, the less they are independent. We agree that the wording was unclear and this sentence was an aside which doesn't contribute to the overall conclusions, so we have removed it.

Page 12, lines 246-247: "This likely relates to the improved response to extratropical diabatic heating in the hindcasts, due to improvements in SPG SST simulation."

- I am not fully clear on how a weakening of the positive AMV-EA 200hpa response, which has a larger magnitude (i.e., weakens further) in the hindcast compared to the historical simulations, relates to an improved response of the model to diabatic heating in the SPG region. While the response weakens further for the hindcast, it remains the wrong sign. Does the further weakening in the hindcast explicitly relate to improvements in the SPG simulation? If so, how is this demonstrated? It could be that in the hindcast and historical simulations both see a weakening of the regression slope with increased lead time purely because the role of the Rossby wave response weakens, not necessarily that another mechanism becomes more important. Additionally, it might be useful to also include 5-95% confidence intervals for the ensemble of historical simulations in Figure 8, as it could be that the central 90% percentile of the slopes overlaps for longer lead time averaging windows, thus indicating no significant difference in the response between the historical and hindcast simulations.

and

Page 15, lines 289-291: "The regression slope of the AMV index on EA 200 hPa geopotential height is calculated for different data sources and rolling window lengths, and a robust reduction in the slope (to negative levels in the case of reanalyses) with increasing rolling window length is found, with the decadal hindcasts performing better than historical simulations."

- Linking to my previous comment on the findings of Figure 8 here, I am not fully clear on how a stronger reduction in the slope of the hindcast with lead time (when the sign is still incorrect) indicates an improved or 'better' response to diabatic heating from the SPG. It might be useful to adjust the language or provide further clarification here (i.e., the weakening in regression slope is directly due to improved SPG response). I think this weakening clearly indicates reduces dominance of the Rossby wave response with increasing rolling window length, but this does not necessarily provide evidence that this is directly caused by improved atmospheric response to diabatic heating in the SPG region.

and

Page 16, lines 335-339: "By adjusting rolling window lengths between 1 and 15 years, it becomes clear that both the high-frequency tropical Rossby wave and low-frequency extratropical diabatic heating response exist in observations and reanalyses, but

deficiencies in the strength of the surface level response in hindcasts means that its role at upper levels increases more slowly with rolling window length than in observations."

- Linking to previous comments, I am not clear on how the weakening of the positive regression slopes with increasing lead time in Figure 8 (red hindcast line) infers an increase in the role of the SPG heating response leading to negative 200 hPa geopotential height anomalies with increasing lead times. How does the weakening of the positive regression slope demonstrate an increase in the surface level diabatic heating response and its role at upper levels, rather than just a weakening of the Rossby wave response?

We have considered these comments, and we agree that the results in Figure 8 are not strong enough to justify this conclusion. We have made significant alterations to the text in Sections 4 and 5 to reflect this. This has bolstered the scientific validity of our study and we appreciate that you pointed this out.

We originally considered including the confidence intervals for the historical simulations but narrowly decided not to in order to reduce clutter, but after this comment we have included them. There is an overlap in the historical and hindcast confidence intervals, but we are keen to point out that this does not imply a lack of significant difference between the two: the confidence intervals of historical (hindcast) slopes do not overlap with the mean hindcast (historical) slopes, indicating that the best estimate hindcast slope is still below the 5% level of historical slopes, and the best estimate historical slope is still above the 95% level of hindcast slopes.

Page 12-13, lines 247-252: "By including all leads—for rolling window length L= 1, leads 1 to 15 are used, for L= 3, leads 1–3, 2–4 up to 13–15 are used, and so on— it is shown that the dependence on rolling window length in the hindcasts is not caused by the inclusion of longer leads with longer rolling windows. Leads 1–L (i.e. the first L summers) are also shown; for L= 3 and L= 7, the slope is significantly more negative than when considering all leads, demonstrating that the improvements are greater at short leads; note that as L tends towards 15, the red squares (lead 1–L) and red circles (all leads) necessarily converge."

- I am not fully clear on how the two different methods of filtering via rolling windows are applied to the hindcast. Particularly, I am not clear on how this is calculated across the different initialisation years for the 20-year hindcast runs. For the first method, if when L=3, leads 1-3, 2-4, up to 13 15 are used, are these slopes calculated for leads 1-3, 2-4, 13-15, across all hindcast initialisation years? Or just for a single initialisation? If they are calculated across hindcast years, how is this done? I am not clear on how either of the methods are

aggregated across initialisation, so would benefit from further detail/explanation here.

We have added the following to the Data and Methods section:

Where regression slopes are computed, they are computed across time. for the hindcasts, each time point is a hindcast from a different initialisation year.

And the following to the caption of Figure 8:

Where the values are for all leads, values are computed individually for each lead, and then averaged.

We agree that is important to clarify this. The 'all leads' point is only in the Figure caption rather than the text as it is difficult to incorporate in the relevant paragraph, and would lack context in the Data and Methods section.

Page 13, Figure 7

- Here, again, it would be useful to include the fixed window (e.g. 1960-2014) over which these responses are calculated. As well, it would be useful to present the 200 hPa geopotential height anomalies with solid white between -2 m and +2 m, to better highlight the centres of action. Additionally, in 7f), it would be useful to clarify in the figure caption the mechanism for calculated the year-to-year differences across initialisation times in the extended hindcast. Does this, a) for a single initialisation, calculate the differences between lead 1 and lead 2, lead 2 and lead 3 etc. or b) across initialisations, calculate the difference between init 1, lead 1, init 2, lead 1, etc., and then aggregate across the 15 leads? It would be useful to have further clarification on precisely how these are calculated. Also, there is likely a type in the figure caption where: "Composite difference in MSLP…" should likely be "Composite differences in 200 hPa geopotential height anomalies".

As with similar figures, we believe that the time period information is sufficiently explained in the Data and Methods section.

We have changed the figure to have a white filled contour centred on zero as suggested.

We agree that more clarity was required on the topic of 'year-on-year' variability. We have added the following to Section 2:

Individual summers include both interannual and longer timescale variability. To understand the role that interannual variability plays, it is useful to filter low frequency variability out. We define year-on-year variability as half of the difference between successive summers. This simple high-pass filter damps the low frequency contribution to interannual variability (Stephenson et al., 2000). This method utilises the fact that for a component of variability with a period much longer than a year, there is very little

variation between consecutive years, meaning that this difference approximately removes this component. For 'true' interannual variability with no autocorrelation between years, the raw difference between years has double the standard deviation, hence need to half the result. When applying this method to hindcasts, successive summers are taken to be at the same lead but with initialisations separated by one year: results are qualitatively the same using differences from the same initialisation separated by a year of lead time.

And referred the reader to Section 2 for more details when year-on-year variability is mentioned in the caption of Figure 7, as well as in the text of Section 4.

Thanks for pointing out the error in the figure caption. We have now fixed this.

Page 13, Lines 254-255: "When considering all leads, the slopes are significantly less than zero for rolling window lengths of 3, 5 and 7 years, whilst the lead 1–L slope is significantly more negative than for all leads for L from 1 to 9."

- Linked to the above point, why would we expect the 1-L rolling window method to better capture the negative EA 200 hPa response to AMV than the alternate rolling window method? Understanding precisely how these are calculated across initialisations (see above) would help to improve understanding of this.

We have added the following after this sentence:

This is consistent with the skilfully predicted SPG SSTs and poorly predicted tropical SSTs demonstrated in Figure 1 (f): the response using ERA5 AMV is negative as expected in response to SPG SSTs, and the skill is likely to be highest at short leads.

We hope this (along with the explanation of the regression slopes across leads) makes things clearer.

Page 14, Figure 8

- Again, it would be useful to define the fixed period over which these regressions are calculated for clarity. Relating to previous comments, it would be useful, either in the main text, or here, to describe in greater detail how the rolling window averaging is applied across different initialisations and precisely how the two different approaches differ.

We have added the sentence:

Hindcast results use all initialisations except where correlated or composited with a variable from reanalysis/observations, in which case only initialisations are used which have target years within the period of the relevant reanalysis/observational dataset.

to the Data and Methods section, after the period used for other datasets has been introduced.

We have also added:

Throughout the study, any averaging of multiple years in the hindcasts is done by taking different leads from the same initialisation, rather than averaging over different initialisations at the same lead.

When introducing the hindcasts in the same section.