We thank both reviewers for their comments. This review response consists of the individual comments already provided, with the addition of line numbers in the revised manuscript where relevant.

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 (L47). 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 (L77):

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 (L390):

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 (L104):

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’ (L118).

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 (L136; **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 (L102):

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 (L200):

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' (L202):

(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 (L326):

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 (L265):

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 (L139):

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 (L149):

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 (L304):

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 (L99):

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 (L80):

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.

<u>Response to Review 2</u>

We thank Reviewer 2 for the time taken to thoroughly review this manuscript, which will allow us to improve it further. We appreciate the constructive suggestions to improve how we have contextualised the study and to improve its structure. Our response to Reviewer 2's comments are in blue below, including changes to the manuscript which we hope will address the concerns raised.

<u>Major Comments</u>

The first paragraph of the introduction contains parts of the motivation, a methodological information, and the overarching research objective. It feels a bit like a repetition of the abstract. The overarching research objective alludes to physical mechanisms that are only explained later. In my opinion, the introduction would benefit from a reordering such that research objectives naturally emerge from introduced concepts or gaps. Currently, you repeat differently phrased "aims of the study" in lines L29-34 and L64, which – up to personal structure preferences – could be circumvented by stating one overarching goal and more detailed ones that hint at the structure of the results. Moreover, while I do appreciate the present conciseness, the AMV (and the word horseshoe pattern) could be introduced earlier or with another sentence.

Thanks for this very helpful comment. We have reordered the introduction, merging the 'aims' part of the first paragraph into the 'aims' part of the final paragraph. We have moved the first sentence of paragraph 3 – introducing AMV and its associated SST anomalies – to the end of paragraph 1 and added the sentence '*This characteristic pattern of AMV SST anomalies is sometimes referred to as the 'horseshoe' pattern'* (L29).

In Section 4, many paragraphs start with "Figure xy shows …". I suggest refining the transitions between these paragraphs such that it becomes clearer what is done next and why. The summary from line L257 is very clear, whereas the discussion of Fig. 7 from L228 that starts with introducing the method of year-to-year variability was not very accessible to me.

Thanks for this comment. Where paragraphs previously started like this, we have added a sentence at the start explaining the scientific motivation behind looking at what the figure shows. The one exception is the second case of a paragraph starting with 'Figure 8', as it follows on from the previous paragraph on Figure 8.

We agree that more clarity was required on the topic of 'year-on-year' filtered variability. have changed the sentence that introduced year-on-year variability to:

Year-on-year variability is computed as half of the difference between successive summers, in order to filter out multi-year variability (see Section 2 for details).

And added the following to Section 2 (L149):

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.

In the “Discussion and Conclusions” section, there is a short (kind of stand-alone) summary at the very end. While this might be a question of style, I suggest that this is reordered or labelled better. Moreover, in the current version the first sentence of the final section reads L264 “The capability of … at capturing spatial patterns of SSTs … was tested.” while the overarching goal stated in the introduction is to assess the L33 “atmospheric response to AMV”. Of course, the atmospheric response is discussed right after, but at first this to me seemed like a disconnect between the stated goals and the summary of the work. Please improve coherence in that regard throughout the paper. Similarly, predictability is the last word of the title (as if it were just a side aspect) but then is a key aspect of the abstract and rest of the paper. The way I understood the manuscript, “Predictability of European summer climate: influence of competing mechanisms related to the AMV” could also be a suitable title.

For the standalone summary, we have decided that the best option is to explicitly label it as such, with ‘*To summarise,*’ added to the start of the first sentence.

We have changed the first sentence in the Discussion and Conclusions section to

The capability of MPI-ESM-LR historical and decadal hindcast simulations at capturing the atmospheric response to Atlantic Multidecadal Variability (AMV)—as well as SST anomalies associated with AMV—was tested’

We agree with your suggested title change, and so we have adopted it with minor alterations:

Predictability of European Summer Climate: The Influence of Competing Mechanisms Related to Atlantic Multidecadal Variability

Can you provide a motivation or reference for why to use exactly 7 years for averaging?

Thanks for pointing out to us that this isn't clear. We use 7 year averages as that is a standard for decadal prediction studies, as it balances the prediction skill benefits of taking multi-year averages and the skill losses with longer lead times. In the manuscript, we now have the following (L78):

Lead year 1–7 means are used except where otherwise stated; this is similar to other studies focused on decadal variability of mid-latitude atmospheric circulation (e.g. Smith et al., 2020).

Please specify the model resolution.

We have now included the atmosphere and ocean grid spacing when we mention that it is the low resolution version of MPI-ESM (L68).

L70 onward: I would have appreciated a (basic) description on how the hindcasts are initialized (like which data is used for initialization) instead of having to go to the indicated references. This would also help to better understand possible limitations.

We have modified the following excerpt:

Members are initialised every year in November from 1960 to 2019 inclusive and are run for a minimum of 10 full calendar years beyond the initialisation date. In addition, the original 16 members have been extended to 20 years after each initialisation date (Düsterhus and Brune, 2024).

It is now (L71):

Members are initialised every year in November from 1960 to 2019 inclusive; the ocean is initialised using temperature and salinity from the EN4 ocean reanalysis (Good et al., 2013) to create a 16 member ensemble Kalman filter assimilation, whilst the atmosphere is nudged to reanalysis, with perturbations applied to the stratosphere to generate 4 extra members per each assimilation member, resulting in a total of 5×16 = 80 members. Further details can be found in Brune and Baehr, (2020). All members are run for at least 10 full calendar years beyond the initialisation date, and the original 16 members have been extended to 20 years (Düsterhus and Brune, 2024).

L89: The anomaly with respect to what precisely?

We have added "*(relative to the climatology over the full period analysed)*" (L103).

I assume that the detrending using the ensemble mean timeseries is applied to both historical and hindcast model runs, is that correct? In L185, you indicate that also the reanalyses data are detrended. Could you specify how this is done? Is this using the trend derived from the model?

The detrending of observations/reanalyses also involves regressing out the free-running ensemble mean GMSST timeseries and we have added (L117):

“This method is intended for use on both observational and model data, and is applied here for all datasets...”

After introducing it. Although it may not seem obvious that this is sensible, it has several advantages. As well as being a ‘best estimate’ of the (globally coherent) externally forced component, it means that the timeseries being regressed out is consistent amongst datasets (as would be the case if regressing out a linear trend); by using ensemble mean GMSST, the method does not depend on model capability in simulating regional differences in trends; by using regression rather than subtracting the timeseries or similar, it is not affected by model errors in climate sensitivity (although Mauritsen et al. (2019) demonstrates that this model has a reasonably accurate climate sensitivity, anyway!).

Figure 1d,e:

- Could the difference between the two panels be explained by increased SST variability in the SPG region from the early historical period towards the hindcast period? Do the historical simulations exhibit a larger AMV signal during the hindcast period only (without the imposed observational constraint present in the hindcast ensemble)? In other words, I was curious to see Fig. 1d repeated for the hindcast period.

  Thanks for this comment; we had considered including the hindcast period composites for the historical simulations, but they are not qualitatively different from the full period for any variable considered. We have added the following sentence in the Data and Methods after introducing the historical simulations (L84):

  Results were not found to be sensitive to the time period of the historical simulations (not shown) and so the full period is used.

  It is also worth noting that Fig. 1 (a) and (b) show observations for the full period of HadISST availability and for the hindcast period respectively, so the observed difference is already included in the manuscript.

- More generally, how do MPI SST trends in the historical ensemble compare to observations in the SST region? Would the model simulate, for instance, a stronger or weaker SPG cooling than observed if not constrained as in the hindcast ensemble?

We agree that this is an interesting question, but as the focus of the manuscript is on the implications for atmospheric prediction, we have decided that this is not within the scope of the manuscript, notwithstanding the relevance of the atmospheric response to regional SST trends. We have added the following at the end of the penultimate paragraph (L413):

The forced trend was removed throughout this study in order to focus on the decadal component of the overall variability, but it is necessary to include the trend in operational prediction of near-term North Atlantic-Europe climate. Recent work has found signal-to-noise errors in the wintertime North Atlantic circulation response to external forcing (Blackport and Fyfe, 2022; Klavans et al., 2021), and so understanding whether this is also an issue during summertime warrants further study, including whether North Atlantic SST trends play a role.

- Are there any systematic model drifts in the hindcast ensemble after initialisation that would not be filtered out by removing the forced trend determined from the historical simulations?

Thanks for this question. The removal of the forced trend does not remove systematic model drifts. In this analysis, we do not at any point rely on differences between different leads, or on full-fields rather than anomalies, and hence the drift doesn't appear directly in the results. However, we recognise that it is important to acknowledge potential errors in simulated variability caused by biases or model drift. We have added the following paragraph to the Discussion and Conclusions section (L397):

The MPI-ESM-LR hindcasts analysed in this study use full-field initialisation (Brune and Baehr, 2020) in the ocean and atmosphere, meaning that they drift from an unbiased (relative to the initialisation data) to a biased state during the model integration. Atmospheric biases typically develop quickly relative to decadal timescales, but noticeable oceanic drifts may persist for several years before stabilising (Hermanson et al., 2018; Polkova et al., 2023). Both biases and drifts interfere with predicted anomalies, but Polkova et al. (2023) found that decadal hindcasts using full-field initialisation (drifts but with reduced biases) as opposed to anomaly-only initialisation (reduces drift but with a larger bias) had higher prediction skill for SPG SSTs. However, they also found that full-field initialised systems with the smallest biases had the highest skill. Additionally, it should be noted that the slower oceanic model drift can be a direct result of integrating the already-developed atmospheric biases, meaning that there is spurious transfer of heat between the ocean and atmosphere (Sanchez-Gomez et al., 2016); this is likely to have an impact on the predicted atmospheric response to oceanic forcing. Whilst we do not investigate biases in this study, the existing literature *demonstrates the necessity for targeted effort to reduce model biases in both the ocean and atmosphere in order to maximise prediction skill.*

Figure 2: From a quick computation, the temporal standard deviation of the JJA mean SLP lies between 2 and 2.5 hPa in ERA5 to the west of the UK. The differences between the AMV+ and AMV- years in Fig. 2 seem comparably low with 0.9 hPa in reanalysis and 0.35 hPa in models. Similarly, most readers likely won't know whether anomalies of 15 m in 200 hPa geopotential height are a lot (on the synoptic scale, for instance, they aren't). Can you contextualize the seasonal signals that you find with natural variability and other studies? To what extent would you argue that your decadal predictions can provide an added benefit to people from the climate impact community (compared to existing studies)?

As we are looking at multiyear averages, the standard deviation is considerably lower – between 0.6 and 0.7 hPa in all data sources. The correlation skill plot demonstrates that a large proportion of the decadal timescale variance is predicted. We have added the following sentence (L198):

The standard deviation of 7 year mean EA MSLP is comparable across all data shown in Figure 2 at between 0.6 and 0.7 hPa, indicating that the HadSLP and ERA5 composite responses explain a considerable proportion of the decadal variability in this region.

And for geopotential height we have added (L241):

The standard deviation of 7 year mean EA 200 hPa geopotential height is between 13 and 14 m for all data in Figure 5 except NOAA 20CR for the period common with the hindcasts, which is around 11 m. This indicates that the reanalysis composite responses to AMV explain a large proportion of the variability in this region.

L142: "The strong SPG signal remains, suggesting that it is highly predictable (consistent with previous studies, e.g. Borchert et al. (2021))" Can't this be easily supported (possibly in the supplement) using spatial correlations as in Fig. 3, for instance, instead of via defining indices and deriving differences of subsets of the time series thereof?

Thanks for this suggestion. We have added this figure in the supplementary and added the following to the relevant paragraph (L181):

Figure S1 shows correlation skill maps for hindcast SSTs relative to ERA5 and HadISST, and indicates that prediction skill is highest in the SPG region.

- L295: Could you further discuss possible reasons for the underestimated ocean-atmosphere coupling? Based on your investigations, do you think that this is related to model resolution, mean-state biases in surface fluxes, vertical stability, or something else? In general, the discussion of possible model biases is rather short.

In response to reviewer 1, we have added the following (bold) to the Discussion and Conclusions section (L389):

The role of positive ocean-atmosphere and eddy feedbacks for the observed summer atmospheric response to AMV warrants further study. ***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 mesocale 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.***

Immediately after this, we have added a paragraph focused on the specific problem of model biases in response to your last comment concerning Figure 1 d,e (see above).

- I encourage the authors to illustrate the way the competing mechanisms act in the reanalysis versus the hindcasts using a figure schematic. This would serve as a concise summary that readers could refer back to while reading.

  In response to this comment, we decided to add a two bullet point summary of the mechanisms in the Discussion and Conclusions section (L355):

  The two opposing mechanisms that link summertime AMV index SSTs to East Atlantic 200 hPa geopotential height can be summarised as follows:

  - A tropical-origin mechanism, in which warm anomalies in the tropical Atlantic lead to rising in the upper troposphere, causing a Rossby wave train from the Caribbean to the extratropical eastern Atlantic, with positive geopotential height anomalies to the west of Great Britain and Ireland. In reanalyses, this dominates on interannual timescales, when tropical SST variability dominates overall North Atlantic SST variability. MPI-ESM-LR hindcasts and historical simulations can simulate this mechanism.

  - An extratropical-origin mechanism, in which diabatic heating associated with warm subpolar gyre SST anomalies leads to a low pressure anomaly in the extratropical eastern Atlantic. In reanalyses, this extends to the upper troposhere and dominates on decadal timescales, when extratropical SST variability dominates. Historical simulations do not capture the surface

response whilst it is captured but severely underestimated in hindcasts, and with the tropical-origin mechanism continuing to dominate in the upper troposphere even on 15-year timescales.

We hope this is useful to summarise the two mechanisms and how they differ in reanalysis compared to MPI-ESM-LR simulations.

Minor Comments

- L25: Studying decadal prediction of summertime temperatures is motivated by "increased risk of extreme heat" on the scale of an individual summer (Rousi et al. 2023). It would be more appropriate to keep the time scale consistent here and, ideally, not refer to extreme surface heat as long as land surface temperatures are not subject of the paper.

  Thanks for this comment. The point we were trying to make is that the circulation changes that the paper covers do have an impact on European surface climate. We agree that it isn't a direct consideration in this study, and so this sentence was not suitable for the first paragraph. We have deleted it and modified a sentence in the second paragraph (changes in bold) to better explain this (L34):

  *The increasing risk of extreme heat in Europe during summer due to anthropogenic influence means that predictions of summer climate are particularly relevant (**Rousi et al., 2023;** Seneviratne et al., 2021; Suarez-Gutierrez et al., 2020)**, especially if dynamically-driven variability exacerbates the externally forced trend**.*

- L11: It reads a bit weird that the "anomaly ... predicts observations".

  We have replaced 'observations' with 'the observed anomaly' (L11).

- L26: You describe "Atlantic Multidecadal Variability" as "associated with low frequency variability of North Atlantic sea surface temperatures" – isn't it rather "defined by" than "associated with" (at least in the present context where only the atmospheric response to the SSTs is studied)?

  We have replaced "*associated with*" with "*characterised by*" (L25).

- L51: Does "components" refer to "atmospheric responses"?

  We have changed this to clarify that we mean the SST components – thanks for pointing out this was unclear. (L46)

- L242: Was this actually "demonstrated" in Figs. 5 and 7, or is it rather "consistent with" them? Maybe I am missing something.

  We have replaced "as demonstrated in Figures…" with "Consistent with Figures…" (L290)

- Figure 1 and others: where it makes sense, including the respective time periods in the subfigure titles could help some readers.

  Thanks for this comment. We have now included this information in figures where correlations between models and observations/reanalysis are computed; otherwise we believe the information in the Data and Methods section is sufficient.

- Figure 4 is very neat. The way you explain it in the text, the arrangement of a), c) and b) makes sense. I assume this arrangement looks better than mirroring panel b) to the left? In the caption, the "(a) (top)" etc. is tautologous.

  Thanks for this comment. We tried moving panel b) to the left in response to this comment, but found that it reduced readability of axis labels, regardless of which pane we put them on. We have removed the duplicated subfigure references in the caption.

- L210: I suggest adding or repeating the relevant literature references here.

  Thanks for this suggestion. We've added relevant references to this sentence (first sentence in Section 4).

- Figure 6 or L214: It seems like there is a simple derivation for this, but could you provide a reference for white noise yielding a L^(-0.5) curve for the ratios at different leads?

  Thanks for bringing this up. The most simple explanation is that under white noise, the multi-year mean is a mean of uncorrelated values, and so the standard deviation is analogous to a standard error calculation. We have clarified this, including an equation and a reference (Wilks, 2011), at the end of the Data and Methods section (L158). We have also added '*(see Section 2)*' to the relevant part of the Figure 6 caption.

- L231 and the discussion section: Adding figure references to statements would significantly facilitate reading through these paragraphs and connecting the key take-aways with your specific analyses.

  Thank you for this suggestion: we have added a reference to Figure 5 where we mention the 7 year z200 response in that paragraph.

  We considered doing so for the Discussion and Conclusions section, but felt that it reduced readability.

  <u>Technical Comments</u>

- Please revise the formatting of figure references.

We are more than happy to consider changing the formatting if you are able to clarify. Figure *i* with subfigure *j* is consistently referred to as Figure i (j).

- Figure 8: The ERA5 line is largely outside of the figure bounds; please consider introducing an axis break or another solution to show the full range of the data (even if the conclusion that it is negative does not rely on it).

  Thanks for this comment. We chose the axis limits to highlight model behaviour, but agree that it would be useful to somehow show the full reanalysis results. We decided that the best solution was to include a reproduction of the figure showing the full curves for the reanalysis in the supplementary, which we now mention in the text and in the caption for Figure 8.