Response to Review 2

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.

Major Comments

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'*.

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:

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:

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.

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:

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)*"

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:

"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:

  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:

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:

  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:

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:

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:

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:

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:

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:

  *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'.

- 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*"

- 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.

- 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…"

- 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. 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.