Williams et al 2026 Reviewer Comments

Title: The Influence of Atlantic Multidecadal Variability on European Summer Climate: Competing Mechanisms and Implications for Prediction

Author(s): Ned C. Williams et al.

MS No.: egusphere-2025-6330

MS type: Research article

General Comments

Overall, this paper makes novel contributions to the field of prediction for the European summer climate by identifying the competing mechanisms driving Atlantic Multidecadal Variability on interannual to decadal timescales. The methodological approaches are well justified, and the figures are mostly clear. I identified several areas which could benefit from further explanation of the methodological approaches used and further clarification of the findings. These changes would help to more clearly communicate the findings to the reader and improve the accessibility and impact of the paper. I hope that the authors will consider addressing the following comments to help achieve this.

Specific Comments

- Pg2, lines 52-53: “The time period for which high quality observational data, particularly for non-surface fields, exists is short relative to the timescale of AMV and therefore…”
    - 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).
- Pg3, lines 70-71: “This study uses an 80 member decadal hindcast ensemble using the low resolution version of the Max-Planck-Institute for Meteorology (MPI) Earth System Model (MPI-ESM-LR; Mauritsen et al., 2019).”
    - 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.
- 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.
- 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.
- 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.
- 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). 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?
- 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.
- 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.
- 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).

- 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.
- 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.
- Pg 9, Figure 4
    - Once again, it would be useful to include the period over which these correlations/regression coefficients are calculated here.
- 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.
- 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?
- 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?

- 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?
- 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.
- 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.
- 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.

- 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".
- 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.
- 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.

- 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.
- 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.
- 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?