We thank both reviewers and the Editor for their substantive reviews of the initial submission. These reviews have served to improve and strengthen the manuscript. Below we respond to the major points raised with supporting results arising from new work shown. We will incorporate these results in revisions to the main paper and Supplement.

In the below, the original reviewers are highlighted in bold. Where both reviewers raised similar points we have collated their review comments and responded to them together.

Reviewer 1

This manuscript addresses a highly relevant and still unresolved question: whether the observed Atlantic Multidecadal Variability (AMV) since 1850 is primarily an internally generated mode of variability or largely an externally forced response. The paper argues that AMV-like variability is largely externally forced and that future North Atlantic SST evolution will therefore be driven mainly by anthropogenic forcing. A major strength of the paper is the use of several SMILEs together with a large CMIP6 ensemble, as well as the inclusion of observational uncertainty through different SST datasets. However, I find that the main conclusion is currently stronger than the supporting analysis. Overall, I think the manuscript has clear potential, but substantial revisions are needed. I therefore recommend major revisions. Specific comments are provided below:

1. Detrending method

The separation of forced and internal variability is central to the paper. The manuscript appropriately acknowledges that AMV estimates are sensitive to methodological choices. However, the actual sensitivity analysis is limited to linear versus quadratic detrending, so the claim that the main conclusions are robust is currently demonstrated only within these two methods. The authors should either test at least one alternative estimate of the forced signal from a different methodological family, for example a regression-based approach using global-mean SST/ near-surface air temperature (SAT), or provide a stronger justification for why such alternatives are not considered appropriate here. At present, citing one paper regarding the limitations of regression-based methods is useful, but not enough on its own to establish methodological robustness.

Second reviewer's comment: l.114-119: As recognized by the authors themselves, AMV index strongly depends on the detrending method. The authors cite a few previous studies showing that, they should at least acknowledge that there are many other ones. In fact, I think a detailed discussion on what the AMV is supposed to represent, what does it mean to remove the warming trend, when one aims at showing the impact of external forcing? Why only the warming? Why only its linear part? I guess this can be related to possible biases in climate sensitivity of the models. Please discuss. The manuscript about the impact of linear detrending of the AMV should also be discussed in my view (Baek et al., 2022).

We thank both reviewers for their inputs here which clearly point to the need for further analysis in this regard. We will now additionally estimate the AMV using a regression-based approach in the revised paper (see Fig 1c below), in which the SST variability associated with externally forced changes is removed by linearly regressing the annual mean SST anomalies at each grid point over a 10-year smoothed global mean SST (60° N - 60° S) time series (Andrews et al., 2020; Robson et al., 2023). Unlike the linear (Fig 1a) and quadratic detrending AMV approaches (Fig 1b), the observed and simulated AMV with a regression-based approach (Fig 1c) shows relatively flat evolution before 1920, though muted troughs and peaks are still visible, they are much weaker. Since 1920, the positive phase in 1920-1960 and negative phase afterward are seen in the observations, albeit considerably weaker than the methods originally chosen, but not any longer in the SMILEs' ensemble mean.

This result is consistent with previous studies, which have shown that the regression over global mean SST anomalies can lead to underestimation of the both observed and stimulated AMV magnitude (Andrews et al., 2020; Frankignoul et al., 2017; Murphy et al., 2017; Robson et al., 2023), given that the global mean SST includes part of the North Atlantic signal, which has then been removed.

In addition, the second reviewer has highlighted the paper by (Baek et al., 2022), who suggested that the role for external forcings on AMV is an artifact of the linear detrending method, and showed that GHGs apparently have little to no influence unique to the North Atlantic, and aerosols exert only modest influences using the CESM1 single forcing large ensemble, with detrending method of regressing the annual global SST field onto standardised AMV indices. The CESM1 single forcing large ensemble stimulation used in Baek et al., (2022) only covers the period from 1920-2004, missing the potential significant cold phase in the early 20th century (which is observational dataset dependent in its magnitude).

Perhaps most importantly though, the regression-based AMV substantially reduces the amplitude of the ensemble-mean AMV signal, the observed AMV generally remains within the ensemble spread across most SMILEs (Fig 2) regardless of the method used to isolate the AMV.

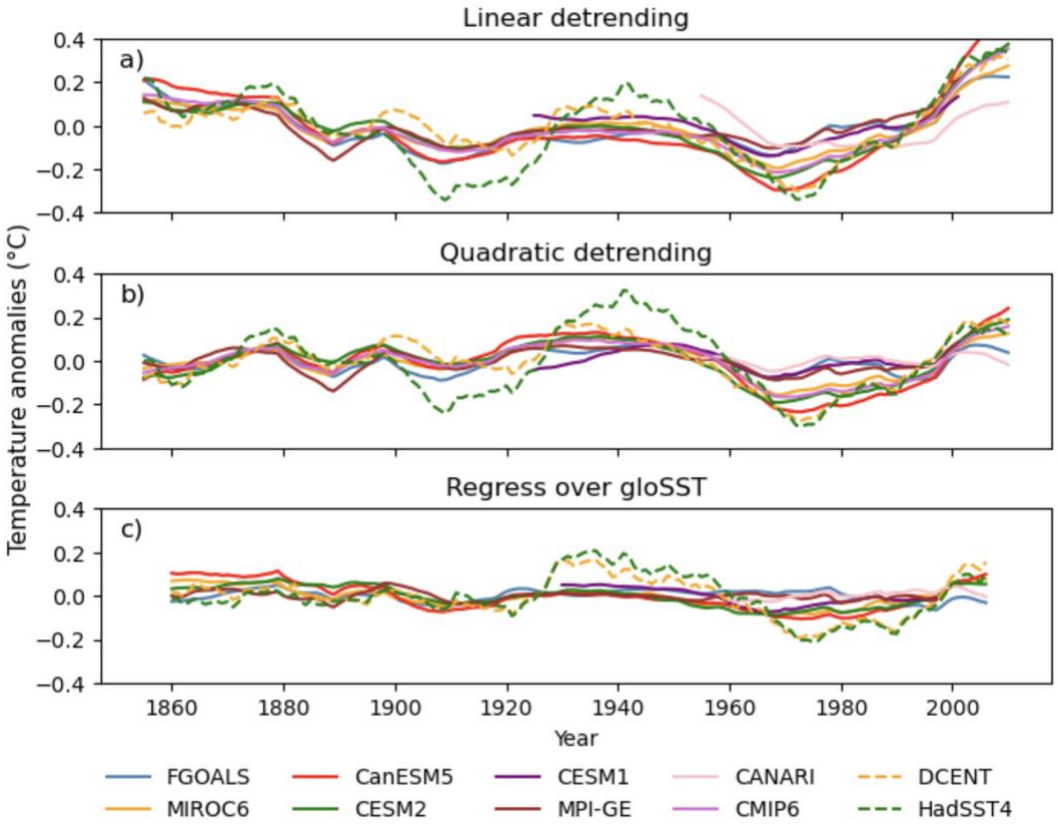


Figure 1. Sensitivity of AMV index to detrending methods. a, AMV indices derived from linear detrending of NASST, compared with HadSST4 (green dashed line) and DCENT (orange dashed line). **b**, same as **a**, using quadratic detrending. **c**, same as a, detrending by regressing over global SST anomalies.

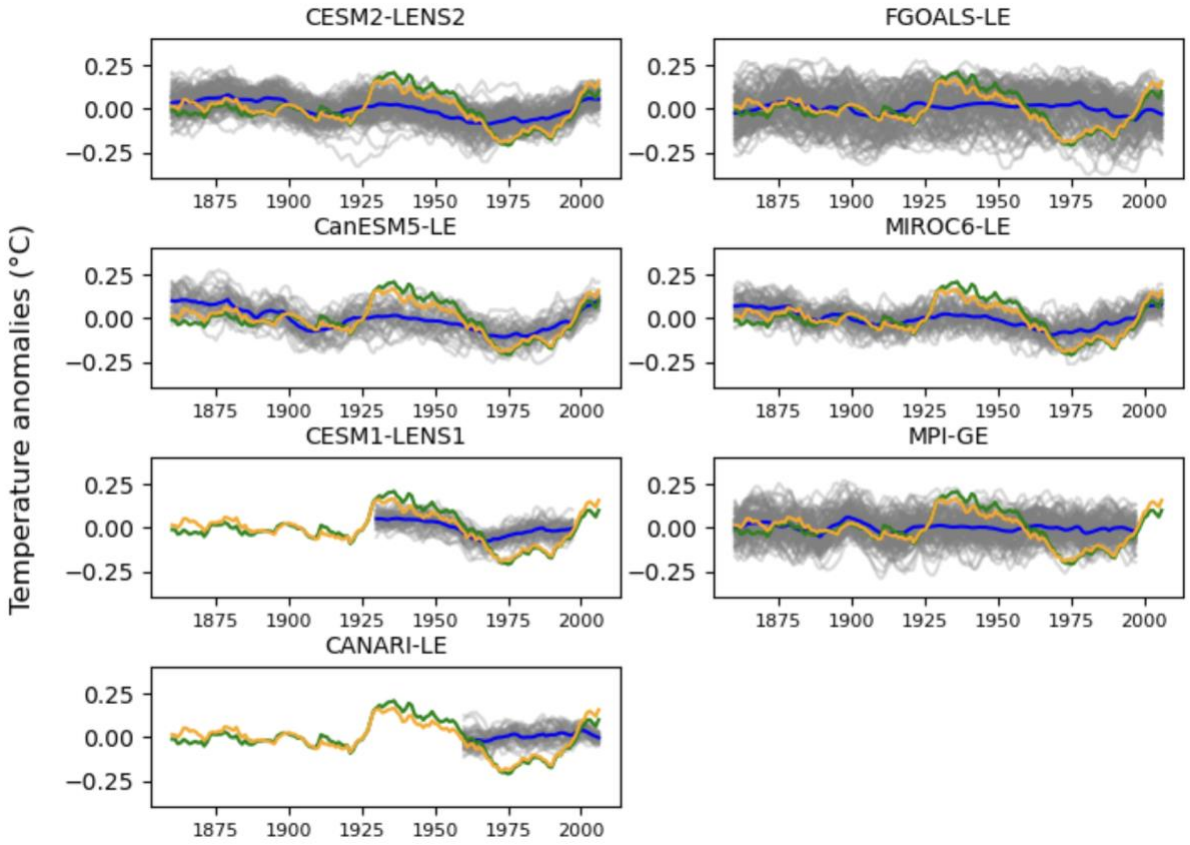


Figure 2. Global SST regressed AMV indices from each SMILEs. Ensemble mean (blue), HadSST (green) and DCENT (orange) are shown.

2. Significance testing

Given that the AMV indices are smoothed with a 10-year running mean, the effective number of independent samples is substantially reduced. The authors should therefore explicitly state how **significance** is assessed and account for serial autocorrelation when interpreting correlation coefficients. An **effective-sample-size correction** such as the **Bretherton lag-1 autocorrelation adjustment** (Bretherton et al., 1999) would be appropriate; alternatively, the robustness of the results could also be checked with a bootstrap or Monte Carlo approach that preserves low-frequency dependence. As it is now it is difficult to judge whether the reported differences in correlation across models and highlighted ensemble members are statistically significant.

The following results have been added to the manuscript with thanks to Maths and stats colleague Rafael de Andrade Morale (added to acknowledgments) for help and guidance.

A lag-1 autocorrelation adjustment is applied (Bretherton et al., 1999; Zhang and Wang, 2013), using:

$N_{eff} = N * (1 - r_1 r_2)/(1 + r_1 r_2)$, where N is the length of the detrended NASST timeseries, and r_1, r_2 is the autocorrelation with the lag of one time step for the observed and simulated time series, respectively.

Then the significance of the correlation between two time series is assessed using a 2-tailed Student t-test with effective-sample-size considered.

$$t_{eff} = r * \sqrt{((N_{eff} - 2)/(1 - r^2))}$$

Using the linear detrended AMV index (smoothed with 10-year rolling mean) for the significance test that applied lag-1 autocorrelation adjustment shows a not a number result (NaN) because when the product of the two autocorrelations is close to a perfect 1 (around 0.9), then the effective sample size goes down to almost zero, and the $N_{eff} - 2$ under square root within the t-value calculation will result in a non-numerical result (Table 1).

Table 1. Comparison of traditional student 2-tailed significance test (p_value) and with lag-1 autocorrelation effective-same-size adjustment (p_eff). The left table is the correlation of the CESM2 ensemble mean with the DCENT, whereas the right table is with HadSST AMV time series.

Correlation of CESM2 EM & DCENT AMV index r (obs&stimulated) = 0.930 N = 156 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.987 r2 (mod lag-1 autocorrelation) =0.996 N_eff = 1.4 p_eff = nan	Correlation of CESM2 EM & HadSST AMV index r = 0.836 N = 156 p-value =0.000 r1 =0.990 r2 =0.996 N_eff = 1.1 p_eff = nan

Alternatively, we use the detrended NASST time series instead. Note that all of the AMV best-correlated members have decreased correlation with respect to the NASST time series than with the smoothed AMV series.

Table 2. Same as Table 1 but use linear detrended NASST time series for both model simulations and observations. In addition to the correlation between ensemble mean and the observations, the correlations between the SMILE's member that is best correlated with the observed AMV is provided.

CESM2: AMV correlation**:** R^2=0.93 with DCENT, R^2=0.84 with HadSST4ca	
CESM2 EM & DCENT NASST r (obs&stimulated) = 0.681 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.584 r2 (mod lag-1 autocorrelation) =0.972 N_eff = 45.5 p_eff = 2.286457072564474e-07	CESM2 EM & HadSST NASST r (obs&stimulated) = 0.689 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.726 r2 (mod lag-1 autocorrelation) =0.972 N_eff = 28.5 p_eff = 4.160661750907124e-05
CESM2 member AMV that best correlated with DCENT r (obs&stimulated) = 0.542 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.584 r2 (mod lag-1 autocorrelation) =0.649 N_eff = 74.3 p_eff = 5.894030363773339e-07	CESM2 member AMV that best correlated with HadSST r (obs&stimulated) = 0.571 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.726 r2 (mod lag-1 autocorrelation) =0.639 N_eff = 60.5 p_eff = 1.693536135549678e-06
FGOALS: R^2=0.65 with DCENT, R^2=0.79 with HadSST4	

Correlation of FGOALS EM & DCENT NASST r (obs&stimulated) = 0.509 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.584 r2 (mod lag-1 autocorrelation) =0.942 N_eff = 47.9 p_eff = 0.00022487085301370335	Correlation of FGOALS EM & HadSST NASST r (obs&stimulated) = 0.666 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.726 r2 (mod lag-1 autocorrelation) =0.942 N_eff = 31.1 p_eff = 4.27252283052848e-05
FGOALS member AMV that best correlated with DCENT r (obs&stimulated) = 0.535 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.584 r2 (mod lag-1 autocorrelation) =0.816 N_eff = 58.5 p_eff = 1.3699495610586254e-05	CESM2 member AMV that best correlated with HadSST r (obs&stimulated) = 0.704 N = 165 p-value =0.000 r1 (obs lag-1 autocorrelation) =0.726 r2 (mod lag-1 autocorrelation) =0.823 N_eff = 41.6 p_eff = 2.2821579359622035e-07

3. Climatology choice

The choice of the 1961–1990 climatology is not justified. The authors should justify this choice and discuss whether the results are sensitive to an alternative standard reference period (e.g., 1981 -2010).

Previous studies have used a range of climatology periods, for example 1981-2010 (Peings et al., 2016) and 1870-2014 (Andrews et al., 2020). For our study we are constrained in this choice by the periods of records of the available family of SMILE simulations. Some do not start before the mid-20th century (e.g., CANARI-LE) and others end in the early 2000s (e.g., MPI-GE and CESM1-LENS).

Here we use observations to demonstrate that the reference period will only affect the NASST timeseries and not the AMV timeseries. While the SST anomalies are shifted, the AMV index after detrending will not be affected by the choice of reference period (they are indistinguishable).

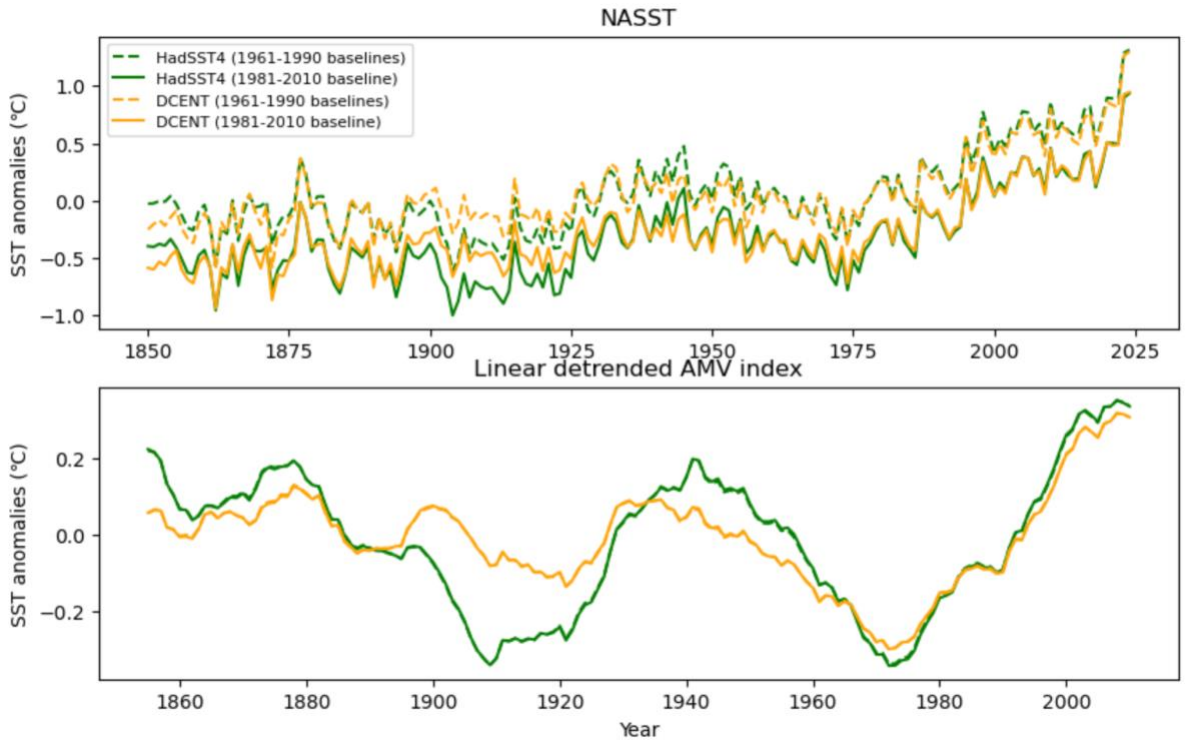


Figure 3. Two climatological choices, 1981-2010 (solid lines) and 1961-1990 (dashed lines) are compared using the observational datasets HadSST (green) and DCENT (orange). The above plot is the NASST anomalies, and the below is the linear NASST detrended over the historical period of 1850-2014.

4. Model resolution

The manuscript should clarify how much model resolution can realistically matter for the chosen diagnostic. Since all SST fields are remapped to a common 5° × 5° grid and then reduced to a basin-mean NASST index, the potential advantages of higher resolution are not directly assessed.

The principal diagnostic used throughout the paper is the basinwide AMV index. As such the choices of aggregation method should have negligible impact upon our results. Nevertheless resolution will affect the ability of models to capture key sub-basin processes such as ocean eddies and western boundary current structure as is well reported in the literature (e.g., Lai et al., 2022). The time and space scales of these processes are several orders of magnitude removed from basin-wide decadal scale processes. Nevertheless, the ability to create and sustain multidecadal variability might plausibly be related to the realism of modelling these processes. Table 3 shows the model resolution.

Table 3. List of SMILE simulations used in this study.

Project	Size	Horizontal resolution (Lat×Lon)	Historical coverage	Historical forcing	Reference
MPI-GE	100	1.8°×1.8°	1850-2005	CMIP5	(Maher et al., 2019)
CESM1-LENS	40	1° ×1°	1920-2005	CMIP5	(Kay et al., 2015)
CESM2-LENS2	50	1° ×1°	1850-2014	CMIP6	(Rodgers et al., 2021)
CANARI-LE	40	~60km	1950-2014	CMIP6	(Williams et al., 2018)
MIROC6-LE	50	1.4° ×1.4°	1850-2014	CMIP6	(Shiogama et al., 2023)
CanESM5	35	2.8° ×2.8°	1850-2014	CMIP6	(Swart et al., 2019)
FGOALS-g3 Super LE	110	2° ×2.25°	1850-2014	CMIP6	(Zhao et al., 2023)

On first impressions it may be argued that the multidecadal variability correlates with the model resolution and ensemble size, given that the two strong multidecadal variability models, FGOALS super LE and MPI-GE have the coarser resolution and the largest ensembles. However, CanESM5 has the lowest resolution among the SMILEs and does not exhibit high intrinsic multidecadal variability. There are many other factors that can affect the model's multidecadal variability and the sample size

of available SMILEs is far too small to make reliable and rigorous inferences here as to any effect of model resolution on AMV behaviour.

5. Spatial Map

A more complete evaluation of the AMV should include its spatial fingerprint, not only time-series correlations of indices. In the literature, the positive phase of AMV is associated with a characteristic Atlantic SST pattern, including warm anomalies across much of the North Atlantic and, in some studies, a North-Atlantic warm / South-Atlantic cool contrast, together with broader SAT responses over surrounding continents (e.g., Bellomo et al., 2018; Deser et al., 2010; Otterå et al., 2010; Ruprich-Robert et al., 2017). I suggest that the authors complement their index-based correlation. This could be done with pattern-correlations between regression maps of detrended Atlantic SST and/or SAT onto their AMV index. This quantification would allow a more physically grounded comparison and could substantially strengthen the paper.

In looking at the literature recommended, there is heterogeneity in the methods used to compute the spatial maps: regression of detrended and smoothed (20-year low-pass filter) NASST on normalised AMV index (Otterå et al., 2010), regression of smoothed (20-year low-pass filter) NASST on normalised AMV index (Bellomo et al., 2018), and regression of smoothed (10-year moving average) NASST on normalised AMV index (Robson et al., 2023). It was not immediately clear to us how these approaches might plausibly help in interpretation of our results and aid the reader in their interpretation.

To try to address the reviewer's concern we decided to look at the **linear decadal trends** over successive warming and cooling phases instead. Here, the AMV time series is divided into two periods of decreasing (1855-1910,1940-1972) and two periods of increasing (1910-1940,1972-2010) AMV index.

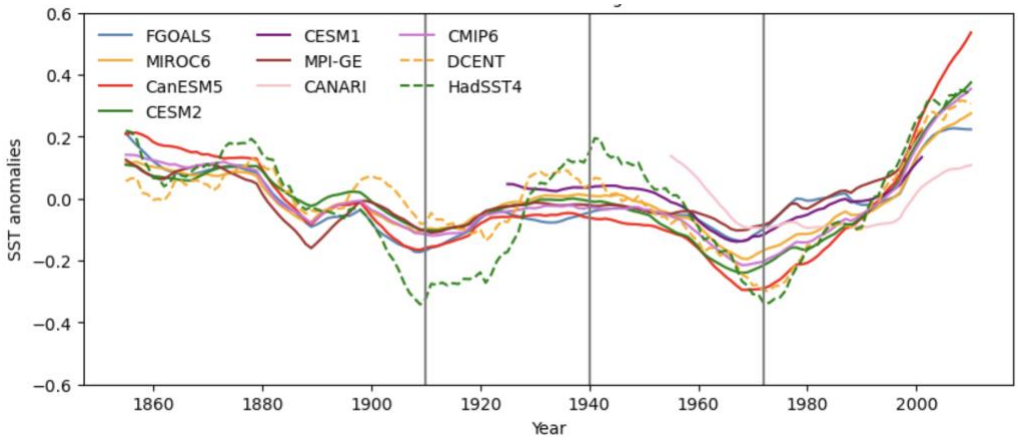


Figure 4. Linear detrended AMV time series from SMILEs and CMIP6 ensemble mean. The grey lines split successive warming and cooling phases.

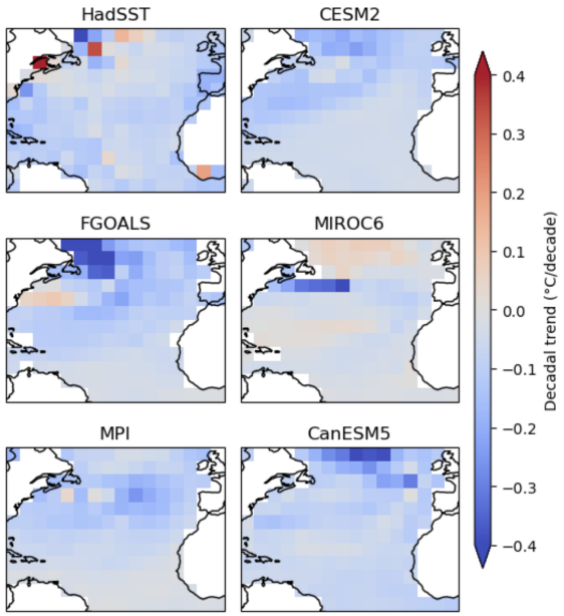
Note:

1. Regridding using the Python package xesmf causes some missing grid point data.
2. CANARI-LE (hist: 1950-2014), MPI-GE (hist: 1850-2005) and CESM1-LENS (hist: 1920-2005) do not have full historical temporal coverage as in other SMILEs, so could not be included in all cases.

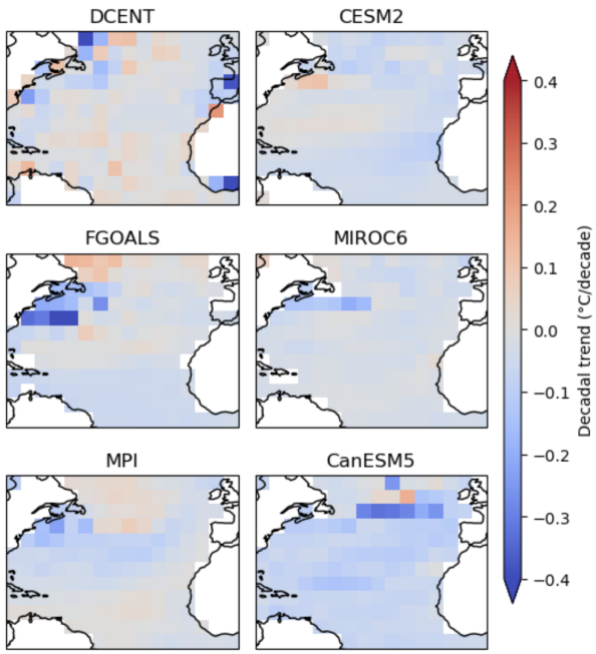
Here, comparison between the observations and the SMILEs' best correlated members are shown in figure 5. It is not entirely clear how to take the spatial analysis further than noting the general basin-wide nature of the patterns in the models and the observations in each period. Any pattern disagreements generally arise in the polar gyre region. Differences between HadSST4 and DCENT, particularly in the two earliest phases, further complicates how to interpret the spatial maps in a robust manner that would add value to our analyses.

Cooling phase 1855-1910

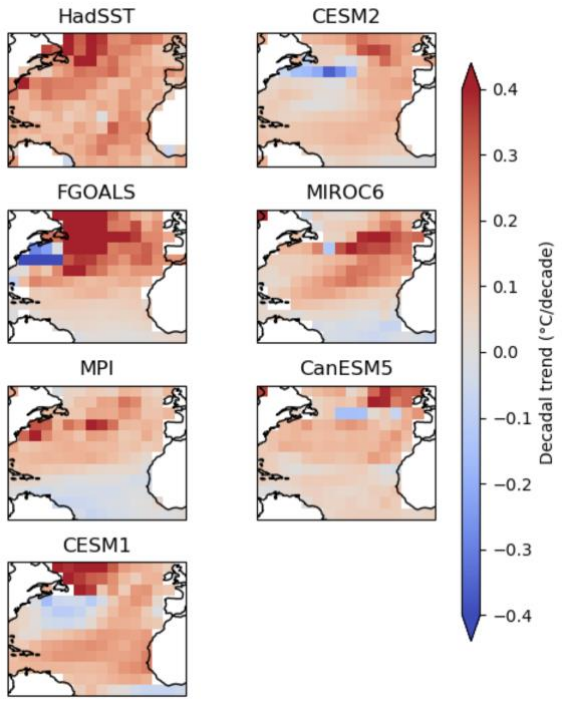
Best-correlated SMILEs' members with HadSST

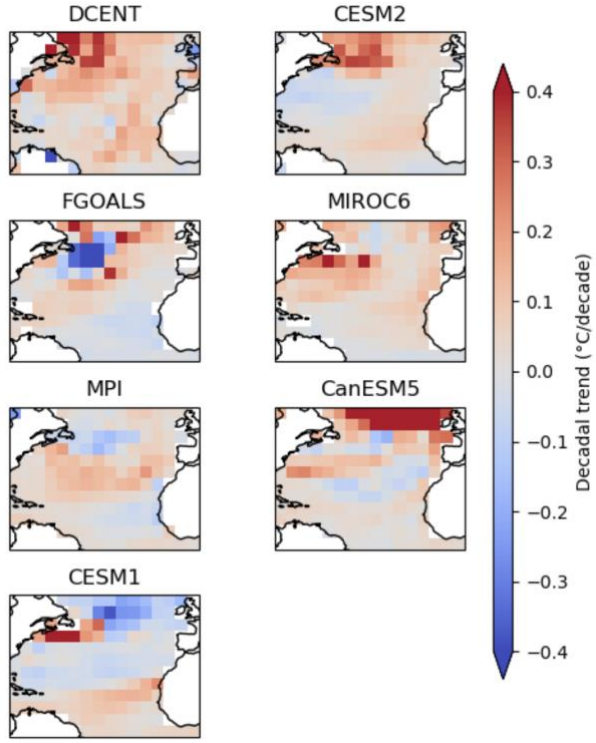


Best-correlated SMILEs' members with DCENT



Warming phase 1910-1940:





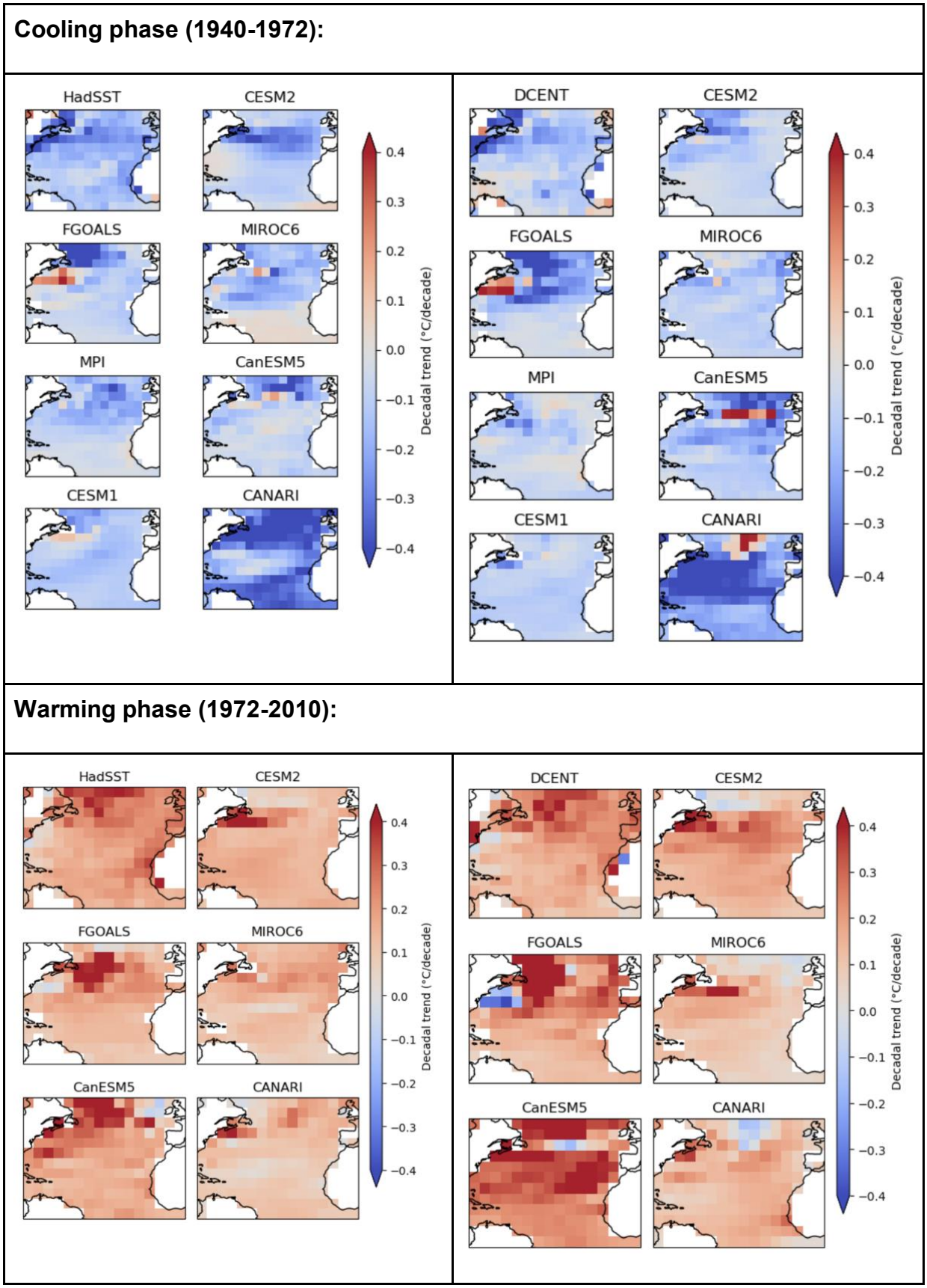


Figure 5. Spatial footprint of AMV's linear decadal changes during successive cold and warm periods. Comparisons are shown between observations, HadSST (left column) and DCENT (right column), and the corresponding SMILEs' members from the linear detrended AMV indices.

Reviewer 2

1. Strong/weak model's multidecadal variability

Distinction among models variability is not clear and not sufficiently illustrated. I am missing some quantification of frequency band as well as amplitude, and a comparison to observations. 'strong multidecadal variability' is too vague to make this a proper conclusion in my view.

Figure 6 shows power spectra differ between each SMILE's detrended annual NASST, and compared to the observations. Most of the models and observations exhibit an energy peak around 80 years. The two models that exhibit strong multidecadal variability, F-GOALS and MPI, show marked decadal variability, with 4 peaks at periods around 20, 30, 40 and 80 years, that are most pronounced in F-GOALS. While no model ensemble tracks completely the observed power spectra: i) there is a large observational uncertainty evident by the substantial offset at multidecadal periods between HadSST4 and DCENT; ii) MPI and FGOALS are most dissimilar. These power spectra results will be added in revisions.

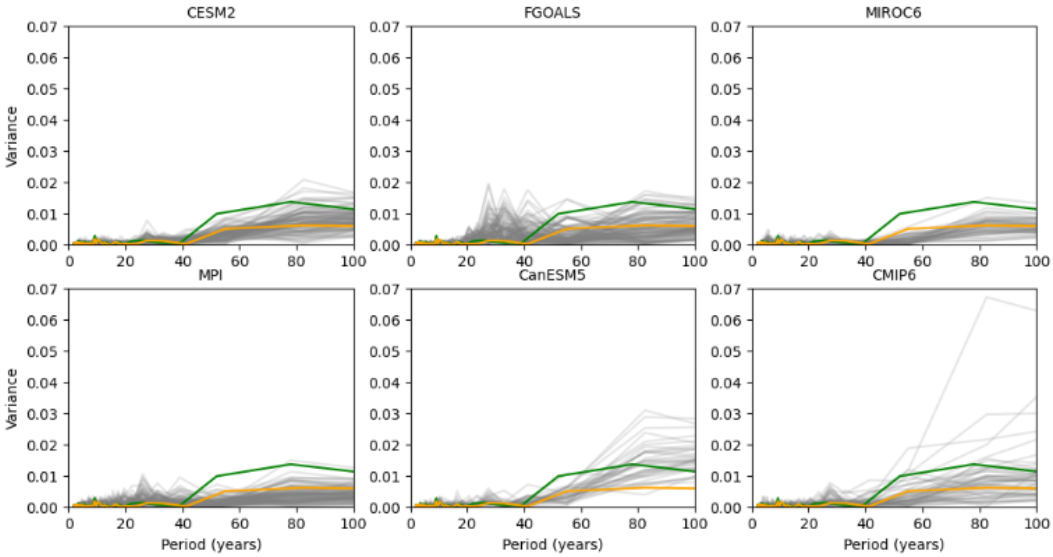


Figure 6. Power spectrum of detrended NASST anomalies from different SMILEs, HadSST (green) and DCENT (orange).

2. SST records biases

The abstract is misleading on the assessment of SST biases vs apparent oscillatory behaviour in the early 20th I don't see where this point is assessed in the paper.

Greater emphasis will be given in the revisions on how the potential biases in SST observation complicates AMV interpretation. The availability of a substantively new and novel SST dataset for the first time in over a decade offers major opportunities for scientific investigation. The SMILEs show more consistent agreement with the DCENT dataset, suggesting that DCENT might be a more physically realistic estimate of the real-world behaviour. This is consistent with emerging literature cited in the manuscript. It would also not be the first time that model-observational discrepancies have eventually turned out to be largely an artefact of residual observational biases (e.g. the tropospheric warming discrepancy in the early 1990s). We will better highlight the value of these new data in revisions.

3. Novelty of the method

SMILEs are not such novel tools, they have been standing for a little decade now (Deser et al 2020). Concluding on a simple ensemble mean is a little bit outdated in my view. How does this approach relate to the current initiative to isolate external forcing from simulations and observations (Wills et al., 2026). Also the difference to multimodel ensemble mean(cf for example IPCC report Fig. 3.40) should be better underlined. Finally, the authors use member selection, which is an interesting use of SMILES. I suggest that they push this approach further (Bonnet et al., 2021).

While SMILEs have stood for a decade, much of the literature has focused on analyses using one or two individual SMILE archives. The novelty of this study does not lie in introducing SMILEs themselves, but rather in conducting a systematic comparative assessment across a broad subset of available SMILEs within a standard methodological framework. This is complemented by the use of two distinct observationally based estimates to explore the impacts of structural uncertainty in both models and observations. These novelty aspects will be strengthened in the revised manuscript.

We acknowledge that relying solely on simple ensemble means can be viewed as insufficient in the context of recent developments in different methods to separate forced response from internal variability (e.g. Wills et al., 2026). Each of the SMILE consists of at least 30 realisations, which is generally considered sufficient to draw a robust estimate of the forced response from the ensemble mean (Frankcombe et al., 2018). We therefore consider the ensemble-mean-based approach, balanced by the single ensemble member analysis to be adequate for supporting the conclusions of this study.

More generally, there is a balance to be attained between methodological complexity and ease of reader understanding. While we recognise the desire of both reviewers for using more complex approaches, their deployment would potentially come at the cost of broad readership understanding. The broad signals that the apparent phasing of AMV appears externally forced and that observational uncertainty matters for interpretation are clear without deploying complex methods that might diminish rather than improve the ability of non-specialists to understand the analysis and its implications.

Specific comments

l.17 the sentence [New insights… behaviour] seems misplaced to me, move to after setting all the SMILE-based conclusions. Anyway, this assertion is not addressed in the text in my view. I would remove this sentence from the conclusion.

Greater emphasis will be placed on how the potential biases in SST observation complicate AMV interpretation.

l.64-66: internal variability is also intensively addressed with control simulations. I think this should be mentioned and discussed here

Added in revisions

l.71: Such kinds of simulations have already been described l. 65-66

Modified accordingly

l.88-89: the diversity of AMV representation emerging from the various datasets is an important of aspect of the paper. Thus I think the datasets should be briefly described. Most importantly since DCENT is a fairly novel and less standard dataset

Added in revisions

I.109: it would not harm to show that masking has a minor effect, especially since the authors partly orientate this manuscript in an analysis of the observations cavevats.

We show below that masking has little impact. A subset of this analysis will be shown in the Supplement and referenced appropriately from the main manuscript.

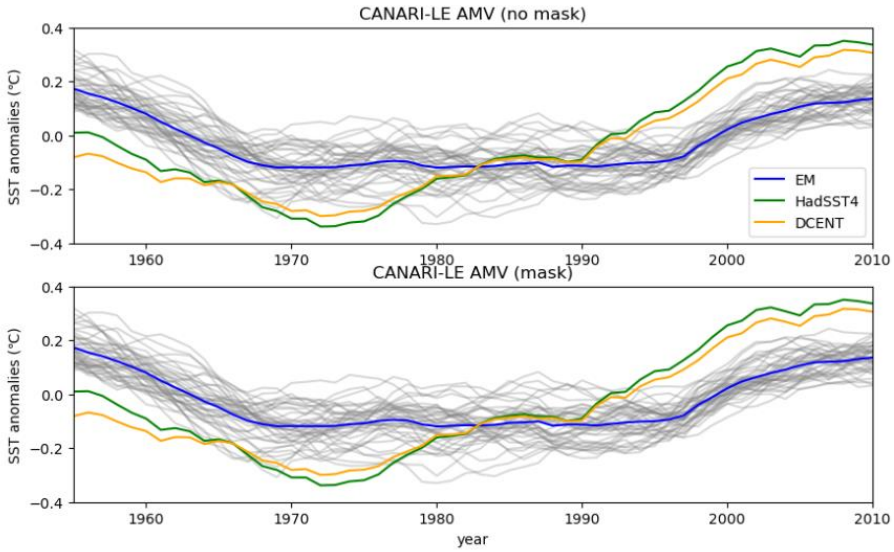


Figure 7. The upper panel shows the AMV indices derived from unmasked simulated CANARI-LE SST, which is used throughout this study. The lower panel shows the corresponding AMV indices after applying spatial and temporal masking consistent with the HadSST dataset.

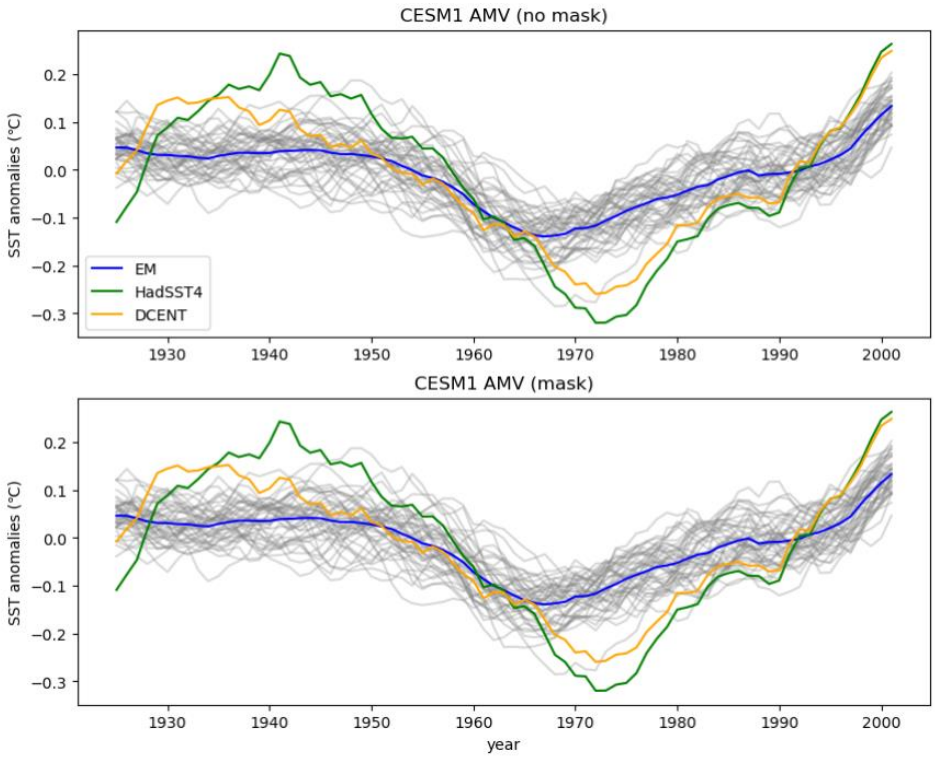


Figure 8. Same as Fig7, but for CESM1-LENS.

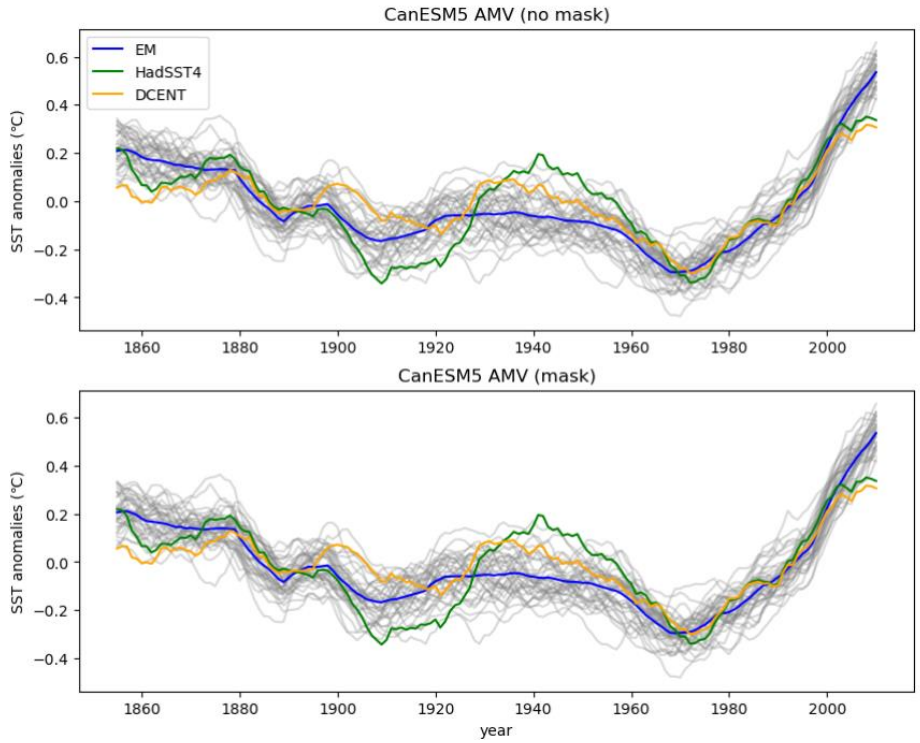


Figure 9. Same as Fig7, but for CanESM5-LE.

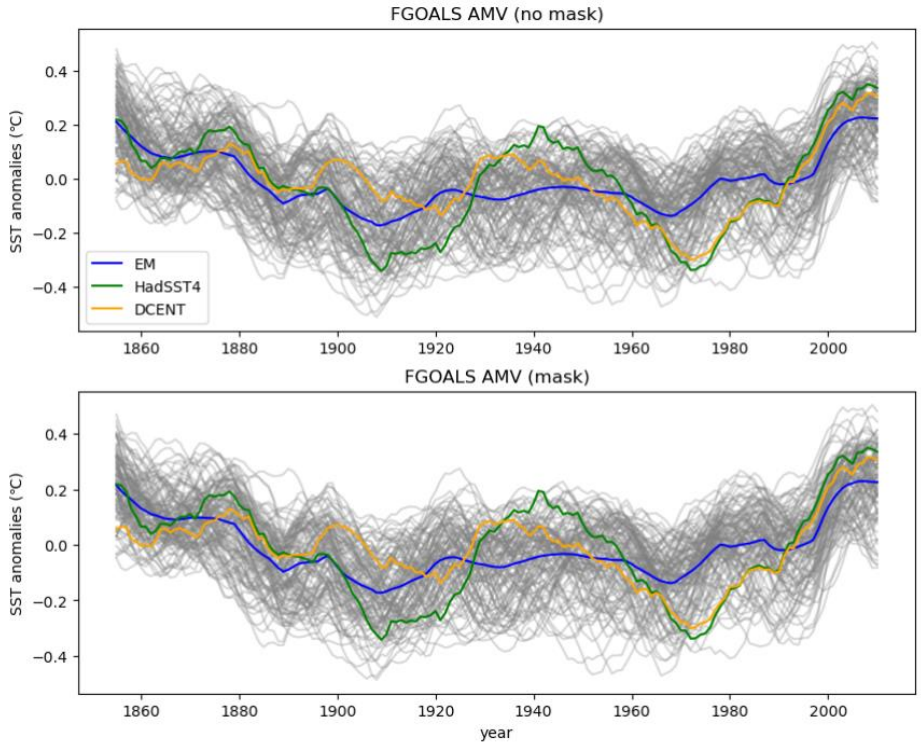


Figure 10. Same as Fig7, but for FGOALs super LE.

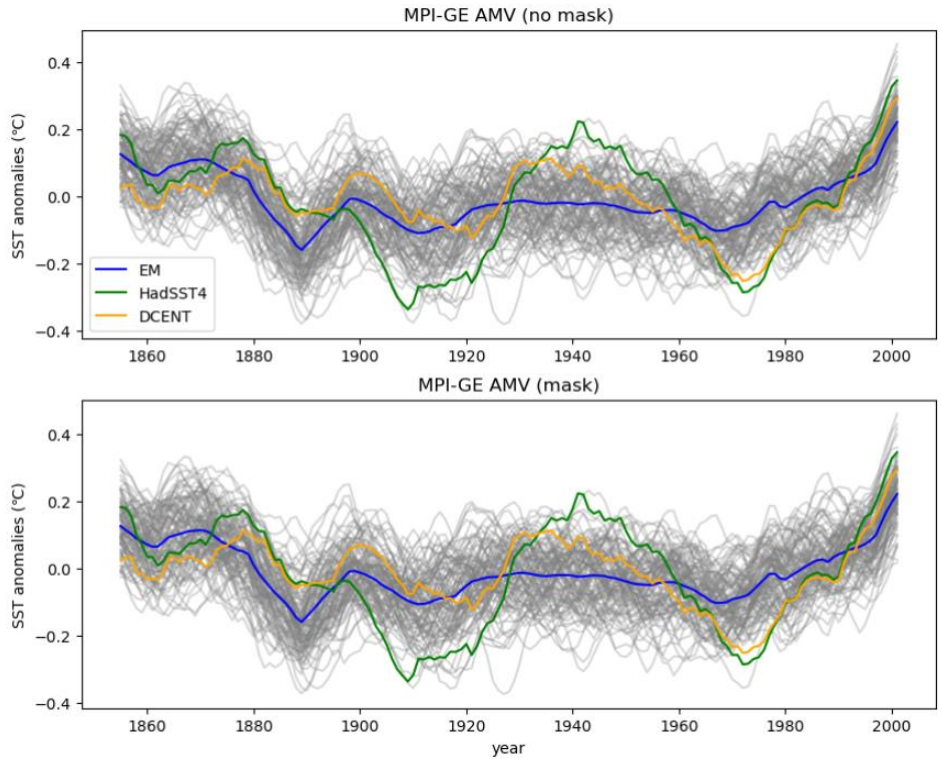


Figure 11. Same as Fig7, but for MPI-GE.

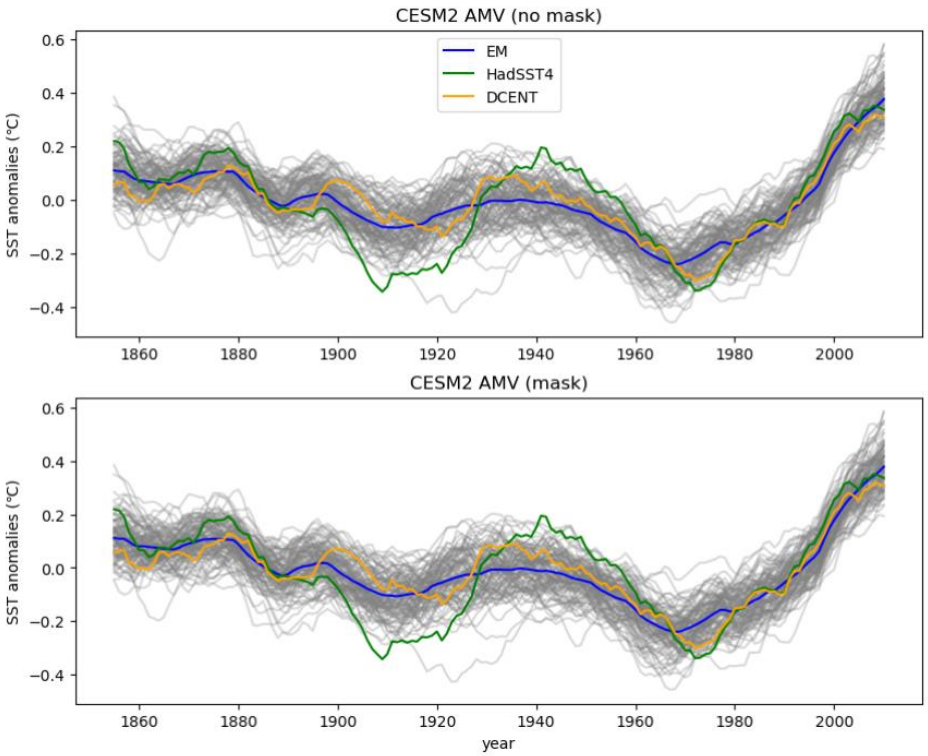


Figure 12. Same as Fig7, but for CESM2-LENS2.

l.114-119: As recognized by the authors themselves, AMV index strongly depends on the detrending method. The authors cite a few previous studies showing that, they should at least acknowledge that there are many other ones. In fact, I think a detailed discussion on what the AMV is supposed to represent, what does it mean to remove the warming trend, when one aims at showing the impact of external forcing? Why only the warming? Why only its linear part? I guess this can be related to possible biases in climate sensitivity of the models. Please discuss.

This manuscript: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022GL097794 about the impact of linear detrending of the AMV should also be discussed in my view.

Addressed in the response to the first reviewer's comments, section 1) Detrending method (page 2). Also, we will address this as part of larger amendments to the text arising from the consideration of the major comments on the same topic.

l.130 : ERSST5 is eliminated because it differs from the other dataset. I don't think this is a valuable reason, specifically since it was originally included to explore data diversity. If the authors decide to eliminate ERSST5 on such wavy argument, I suggest not mentioning it at all.

We will improve our discussion in this regard and reconsider inclusion in the revised manuscript.

l.145 "on the existence of 2 ensemble of models, with limited decadal variability and substantial variability: this distinction should be shown and better explored. What does "limited" and "substantial" mean here? compared to what? Fig. 2 tends to suggest that the difference lies in the frequency more than the amplitude of the variability. Please clarify.

The Spectral analysis shown in response to the major comments will be included to address this.

l.145: "in their ensembles": do you mean ensemble mean? Or each members?

The entire ensemble.

l.148-150: significance of correlations and all statistical estimates is crudely missing. Addressed in the response to the first reviewer's comments on this issue, in Section 2) significance testing (page 5).

Fig. 2: units are missing in the top panels. This is very important to realize whether the models reproduce the observed AMV magnitude.

Added.

l.166: I don't agree that 0.84 is "considerably lower" than 0.79.

Will be amended in redraft.

l.191: "This suggests": I am not sure I agree with the logical link. This "could suggest" but this could also be by chance?

We will amend to clarify what we intended here.

The paragraphs ends on suggesting that because DCENT correlates better with a single member of a climate model, it represents the reality more accurately. Isn't the argument twisted here?

We agree we could have been much clearer what we intended here and this will be addressed as part of broader edits to address the concerns raised around the assessment of the structural uncertainties in observational datasets raised in the major comments.

l.241-244 the sentence is not clear: what role do the models with limited intrinsic multidecadal variability play in the conclusion introduced by "indicating"?

We will clarify intent in the redraft of this passage.

l.253: I don't understand why models with stronger multidecadal variability might be expected to better capture the observed variability? Perhaps you need to quantify this "stronger" (in terms of amplitude, multidecadal variance / total variance, for example) and compare the frequency band with the one detected in observations?

We will better state our argumentation here. But, at a basic level we do not expect a SMILE ensemble that captures the real-world physical AMV if it were a naturally occurring feature of the climate system if it is not initialised from the observed state. Rather we would expect for a sufficiently large ensemble the majority of ensemble members to be out of phase with the real world but a small number, by chance, to be

close to in phase. Therefore most ensemble members would display poorer time series agreement being out of phase whereas a subset that is (broadly) in phase might be expected to show markedly improved time series agreement. But this is not the case. We will make our rationale and argument clearer in revisions.

l.266-267: I agree that the AMOC modulates the AMV. Yet in the argument presented by the authors, the AMOC can be modulated by external forcings, then modulating the AMV. I agree with this but I don't understand how this relates to the general topic of the paragraph "discrepancies between observations and projections".

Statement removed.

References

Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., Burke, E., Dittus, A. J., Florek, P., Gray, L. J., Haddad, S., Hardiman, S. C., Hermanson, L., Hodson, D., Hogan, E., Jones, G. S., Knight, J. R., Kuhlbrodt, T., Misios, S., Mizielinski, M. S., Ringer, M. A., Robson, J., and Sutton, R. T.: Historical Simulations With HadGEM3-GC3.1 for CMIP6, J. Adv. Model. Earth Syst., 12, e2019MS001995, https://doi.org/10.1029/2019MS001995, 2020.

Baek, S. H., Kushnir, Y., Ting, M., Smerdon, J. E., and Lora, J. M.: Regional Signatures of Forced North Atlantic SST Variability: A Limited Role for Aerosols and Greenhouse Gases, Geophys. Res. Lett., 49, e2022GL097794, https://doi.org/10.1029/2022GL097794, 2022.

Bonnet, R., Swingedouw, D., Gastineau, G., Boucher, O., Deshayes, J., Hourdin, F., Mignot, J., Servonnat, J., and Sima, A.: Increased risk of near term global warming due to a recent AMOC weakening, Nat. Commun., 12, 6108, https://doi.org/10.1038/s41467-021-26370-0, 2021.

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., and Bladé, I.: The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field, J. Clim., 12, 1990–2009, https://doi.org/10.1175/1520-0442(1999)012%3C1990:TENOSD%3E2.0.CO;2, 1999.

Frankcombe, L. M., England, M. H., Kajtar, J. B., Mann, M. E., and Steinman, B. A.: On the Choice of Ensemble Mean for Estimating the Forced Signal in the Presence of Internal Variability, J. Clim., 31, 5681–5693, https://doi.org/10.1175/JCLI-D-17-0662.1, 2018.

Lai, W. K. M., Robson, J. I., Wilcox, L. J., and Dunstone, N.: Mechanisms of Internal Atlantic Multidecadal Variability in HadGEM3-GC3.1 at Two Different Resolutions, J. Clim., 35, 1365–1383, https://doi.org/10.1175/JCLI-D-21-0281.1, 2022.

Peings, Y., Simpkins, G., and Magnusdottir, G.: Multidecadal fluctuations of the North Atlantic Ocean and feedback on the winter climate in CMIP5 control simulations, J. Geophys. Res. Atmospheres, 121, 2571–2592, https://doi.org/10.1002/2015JD024107, 2016.

Robson, J., Sutton, R., Menary, M. B., and Lai, M. W. K.: Contrasting internally and externally generated Atlantic Multidecadal Variability and the role for AMOC in CMIP6 historical simulations, Philos. Trans. R. Soc. Math. Phys. Eng. Sci., 381, 20220194, https://doi.org/10.1098/rsta.2022.0194, 2023.

Wills, R. C. J., Deser, C., McKinnon, K. A., Phillips, A., Po-Chedley, S., Sippel, S., Merrifield, A. L., Bône, C., Bonfils, C., Camps-Valls, G., Cropper, S., Connolly, C., Duan, S., Durand, H., Feigin, A., Fernandez, M. A., Gastineau, G., Gavrilov, A., Gordon, E., Günther, M., Höver, M., Kravtsov, S., Kuo, Y.-N., Lien, J., Madakumbura, G. D., Mankovich, N., Newman, M., Rader, J., Shi, J.-R., Shin, S.-I., and Varando, G.:

Forced Component Estimation Statistical Method Intercomparison Project (ForceSMIP), https://doi.org/10.1175/JCLI-D-25-0326.1, 2026.