Referee 2:

The authors present an analysis of the internal variability (ICV) in annual and seasonal mean near-surface air temperature over Europe using different data sources, timescales, and methods, showing basic patterns of variability (more in the northeast and during the winter, e.g.) and its change under anthropogenic forcings (less in most seasons and most regions). It's important to wrangle the constantly increasing menagerie of different climate data sources into coherent summaries of existing conditions and future changes; I thus thank the authors for their work.

As it stands, however, I believe too many things are varied between the different sources of information about ICV, making it difficult to understand whether a given finding is because of the data source, the method used to characterize ICV, the timeframe of the data used, or some combination of the above. I also have a few methodological disagreements, or at least requests for additional clarifications. More broadly, I think the goals of the paper could be clearer - especially how they imagine these results to be used to "benchmark" regional climate simulations.

As such, I recommend Major Revisions before publication.

Thank you for all the comments. We have addressed all the comments, please find the relies below. We have also performed more calculations and analyses, to enable better intercomparison of the results for individual methods.

Here are abbreviations of the methods used in the study and their description:

For CMIP6 preindustrial simulations we have used two different approaches:

Pidata 30yrs-STD: calculation of standard deviation (STD) for non-overlapping 30-years period timeseries within the period 1850–2349. For each of the models we have 16 STD values on each of seasonal and annual time scales.

Pidata 500yrs-STD: calculation of STD for 500-years period timeseries (for the period 1850–2349), which yields one STD value for each of the model on each of seasonal and annual time scales.

The method from the paper of Hawkins and Sutton (2009) has now been applied not only for the CMIP6 models, but also for observations and SMILE simulations:

HS09 method (Hawkins and Sutton (2009)**):** Seasonal and annual anomalies were calculated during the study periods. These anomaly time series were then smoothed by fitting a fourth-order polynomial using ordinary least-squares regression. Subsequently, the smoothed anomalies were subtracted from the original anomalies to obtain the residual component, denoted as Epsilon. Finally, the ICV was quantified as the standard deviation of the Epsilon.

For SMILE simulations, we also applied the "usual" approach published in recent literature (e.g., Martel et al., 2018; Maher et al., 2020; Lehner et al., 2020), which we call

SMILE TYP method: Seasonal and annual anomalies were first computed relative to the long-term mean during the study periods for each SMILE member. Subsequently, the ensemble mean anomaly, representing the forced signal, was calculated across all 100 ensemble members and removed from each member's anomaly. This procedure isolates the internally generated variability by eliminating the influence of the forced distribution shift. The ICV estimate was then quantified as the standard deviation of these residual anomalies across the 100 ensemble members.

Finally, the method based on linear regression (previously called "attribution analysis") has been applied not only to observations, but also to CMIP6 and SMILE simulations. We now use the abbreviation "REG method":

REG method: The temperature time series at individual grid points for the period 1901-2010 were processed using multiple linear regression, with predictors pertaining to greenhouse gas concentration, volcanic aerosol optical depth, and solar activity. The residuals were then aggregated to seasonal and annual means, and the ICV was estimated as standard deviation of the aggregated time series.

Major points

Temporal comparability between observations and models

Statistics on the observations are shown between 1910-2001, while the "historical" climate model output is shown as 1971-2000. Since ICV is not constant in time (as the paper argues), these two metrics are therefore not directly comparable. If the goal is to benchmark GCM-derived ICV estimates, then the same period should likely be used. If the goal is to show the constancy (or change) in ICV in observations, then perhaps showing different time periods, or trends, or some other metric like that may be useful.

Reply: We have performed additional calculations to enable easier comparison of the results of different methods. We have applied the same 30 years period (1971-2000) for HS09 to CMIP6, SMILE and observations. We have applied as well 110 years period (1901-2010) for the regression method to those three datasets. We have not investigated the trends of ICV within the observed dataset. Regarding the temporal changes in ICV, we only analyze the changes between historical and future scenario time periods within the CMIP6 and SMILE framework.

Comparability between different sets of information about ICV

The study uses a different way of estimating ICV for each different data source – standard deviation across residuals after a linear detrending for PI control runs, standard deviation across residuals after a quartic polynomial for CMIP6 historical / SSP runs, standard deviation of residuals vs. the ensemble mean for the MPI large ensemble, and standard deviation (I believe, though it's not clearly specified in section 3.4) around residuals of a linear model depending on GHGs, aerosols, and solar insolation for observations. I recognize that part of this is due to data limitations – residuals around an ensemble mean are of course only possible one has a large ensemble.

However, in many cases, overlapping methods/timeframes/etc. are possible between data products, and are necessary to interpret the comparisons between data sources shown in Figures 4-9.

In particular, (1) the comparison between observed and modeled ICVs over the historic period is likely affected by the different forms of the ICV calculation, as is (2) the comparison between CMIP6 ICVs and the MPI LE ICVs.

1) A quartic polynomial will produce a different central estimate than a linear fit. What does ICV in the observations look like with the HS09 method instead? You could even ask, what does ICV in the models look like with the linear GHG/etc. fit instead? – Note that the CMIP6 historical forcing does include GHGs, aerosols, and solar insolation changes. The aerosol changes in particular in the historical scenario are perhaps not always very realistic, but in theory one could adopt the same linear model from the observed results in the CMIP6 models as well.

2) The HS09 approach has fallen out of favor lately since Large Ensembles now allow characterizing ICV without having to assume a functional form of the "mean" signal that the ICV is the residual of. Since the paper claims the adoption of this method for single runs of CMIP6 models as a major methodological advance, it would be good to diagnose how well this method actually captures the ICV of a model. The MPI LE (or other LEs, see below) allows one to do that, by showing how different ICVs look like when calculated using the HS09 method separately on each ensemble member vs. calculated as the standard deviation of the residuals vs. the ensemble mean.

In short, one could do overlaps in methodology.

- HS09 for observations, CMIP6, and the MPI-LE
- (less important, due to imperfections in the CMIP6 forcing datasets, but possibly interesting: the linear GHG/aerosol/etc.-based model for observations, CMIP6, and the MPI-LE over the historical period)

That would provide apples-to-apples comparisons.

Reply: We appreciate the suggestions. As we mentioned earlier, we have applied HS09 to CMIP6, SMILE and observations. We also applied linear regression method to CMIP6 and SMILE as an addition to the observations.

Also, we have added some discussion about the differences between the methods. For example: “The methodological advantage of HS09 is that it does not require any knowledge about the external forcings behaviour, it simply assumes that their influence can be estimated by the 4th order polynomial trend generated for individual datasets, and it can therefore be applied to any timeseries that is expected to include both internal and low-frequency externally forced variability. On the other hand, the REG method has the advantage of directly considering a specific set of external forcings (including faster-variable ones such as solar or volcanic forcings), and similar to HS09, it is computationally cheap and can be applied to any timeseries of interest. A disadvantage of the REG method (at least in its specific form applied here) is that it only takes into account the linear response to the external forcings, potentially overestimating the ICV (leaving the non-linear component of the externally forced variability within the ICV estimate). The main disadvantage of the SMILE TYP method is that it requires a large set of model simulations, which are not currently available for most of the CMIP6 models. Similarly, the Picontrol-based methods require computationally demanding datasets, that are not generally available. However, both SMILE and Picontrol allow, by their construction, to distinguish explicitly the externally and internally forced components of the climate variability. On the other hand, they rely solely on climate models, which differ from the real world. The REG and HS09 methods allow for a direct comparison between models and observations.”

General suitability of the ICV calculation methods.

Given that the ICV estimate in Figure 4-9 using the full 1971-2100 CMIP timeseries is generally larger than both the 1971-2000 and the 2070-2099 estimates, I worry that the methods used to estimate ICV are capturing spurious signals or part of the mean trend instead of the internal variability desired.

(are the mean change signals in this data truly linear – in the case of the observations - or quartic – in the case of the CMIP models?)

Reply: We thank the reviewer for this comment.

According to the updated results, we have applied HS09 to CMIP6, SMILE and observations for 30 years (1971-2000) in the main manuscript. However, we additionally applied HS09 to individual CMIP6 models for 130 years (1971-2100) in the supplement along with the historical (1971-2000) and future (2070-2099) periods to examine how individual models estimate ICV in historical, future, and longer time periods. We acknowledge that the larger ICV estimated from the full 1971–2100 period compared with the two 30-year periods could suggest that some low-frequency variability or residual forced signal remains in the detrended series (Lehner et al 2020, https://doi.org/10.5194/esd-11-491-2020). This is a recognized limitation of polynomial-based uncertainty partitioning methods and has been discussed in subsequent evaluations of the Hawkins and Sutton framework. The original HS09 approach provides a reasonably good estimate at large spatial scales but can suffer methodological biases at local/regional scales because internal variability and forced response are not perfectly separated (Lehner et al 2020). Nevertheless, the approach remains widely used for separating forced and internal components in CMIP simulations and was applied consistently across all datasets in this study.

Originally, our study applied the linear regression method to the observational datasets only and the quartic method (Hawkins and Sutton 2009) to CMIP6 only. However, as we mentioned earlier, based on both reviewers' suggestions, we have applied HS09 not only to CMIP6 but also to SMILE and observations. Similarly, we have applied linear regression method not only to observations but also to CMIP6 and SMILE.

Moreover, the linear and quartic functions were not selected because the underlying climate signals are assumed to be truly linear or quartic. Rather, they follow two different published approaches for estimating ICV and we wanted to compare ICV estimates obtained from different methods and datasets. Therefore, the use of different trend representations reflects the original methodological frameworks being evaluated rather than an assumption that the true climate signal is exactly linear or exactly quartic. And the objective of this study is to compare ICV estimates obtained from these different established methodologies. We added discussion of these issues into the last section of the manuscript.

Large Ensembles

The paper correctly notes that using the MPI LE is just one model's estimate of ICV, and subject to the same type of model differences as one can see within the CMIP6 models used here. One could use more models to test this explicitly – for example, the Multi-Model Large Ensemble Archive (MMLEA, Maher et al. 2025) is a set of pre-processed monthly 1-deg large ensemble runs that include the MPI LE but also include the other large ensembles of the CMIP6 generation. It would be relatively straightforward to include those other models in the analysis, which would give an idea of how model uncertainty affects the ICV calculation using the large ensemble methodology.

Reply: Thank you for this suggestion. In the present study, we focus only on MPI-ESM SMILE and we have acknowledged that in the conclusion as a limitation of the study. The reason for the choice is also stated in the text: "we use a single-model initial-condition large ensemble (SMILE) of the MPI-ESM model comprising 100 members (Maher et al., 2019), which, to the best of our knowledge, is the largest ensemble currently available." The comparison with other available large ensembles would be burdened with the uncertainty related to the different number of ensemble members.

Clearer description of what each data source is providing

In the discussion, especially after L342, I'm still a little lost on how the methods relate to the goal of "[being…] used as benchmarks for assessing regional climate simulations over Europe." Perhaps it would help to define some of these terms (especially what you mean by benchmark) and provide an illustrative example of each. As you say, each source of data provides different information; but this is not clearly linked in this paragraph to which circumstances, or for which specific purposes that information could be useful.

Reply: Thank you for pointing this out. We have explained the benchmarking purposes in the introduction: "Conceptually, ICV represents the intrinsic 'noise' of the climate system. Therefore, the ICV estimate has been widely used as a "benchmark" for assessing the climate model outputs. Here, the term 'benchmark' is not used in the sense of directly evaluating model skill in reproducing ICV. Rather, the ICV estimate is employed as a reference scale against which model performance and projected changes are assessed. Model biases smaller than the estimated ICV can be interpreted as being within the range of internal variability, and thus indicate an adequate simulation of the characteristics of interest. Similarly, projected changes that do not exceed the ICV may be considered indistinguishable from internal variability, and hence not statistically robust. While ICV represents internal fluctuations, external forcing can modulate both the mean climate and the characteristics of variability itself. Understanding how ICV varies across regions, seasons, and climatic conditions is therefore crucial for detecting externally forced climate change and for assessing future climate risks. However, ICV cannot be directly observed and must be inferred from observations or model simulations. Different methods and data sources can yield substantially different estimates of ICV magnitude."

Interpretation of PI control results.

I'm not sure why PI control results would be a better benchmark of regional climate simulations than observations over the same time period as the forcings used in those regional climate simulations. I at least would prefer a regional climate simulation to replicate the latter rather than the former. Do you mean instead that they should be used as a comparison baseline climate without anthropogenic climate change?

I'm also not sure what you mean by the PI control simulations having an "advantage" over the others due to constant forcings – whether we want constant PI forcings or not depends on the question we are asking. (and if the goal is comparisons to regional climate simulations, then I have the same question as above, that I'm not sure why regional climate simulations forced with historical or future forcings should be benchmarked against a different set of models with PI forcings instead).

Finally, I'm not sure what is meant by saying the "ensemble mean estimate encompasses the range of the other methods" (e.g., L18, L345) (rather, is it that the PI mean is within the ICVs estimated using the other data sources / methods? Are you implying that we don't think ICV has changed or will change, if it's not differentiable from the PI control estimate?)

Reply: Thank you for these comments. Based on the new results, all interpretations have been changed, including the one stating that Pidata would be a better for benchmarking or an advantage compared to the other methods. The new interpretations have been clarified and addressed.

The sentences in the old lines (L18, L345) have been removed based on the new results.

Regarding the advantages of Picontrol simulations, we have added some discussion on this point: "Furthermore, the Pidata estimates are generally the largest among all methods considered. This likely reflects the combination of longer simulation periods and the absence of external forcings in the Pidata experiments. The extended records provide a more robust characterization of ICV, while the absence of external forcings ensures that the estimated variability is solely internally generated."

Minor points

Trend maps

Could you please display the differences between 1971-2000 and 2070-2099 (in the CMIP6 / LE results) as change maps as well? I personally find it hard to read subtle color differences between, say, the second and third rows of Figure 3.

Reply: We appreciate this comment. We have added the map of differences between future and historical period in Fig. 4.

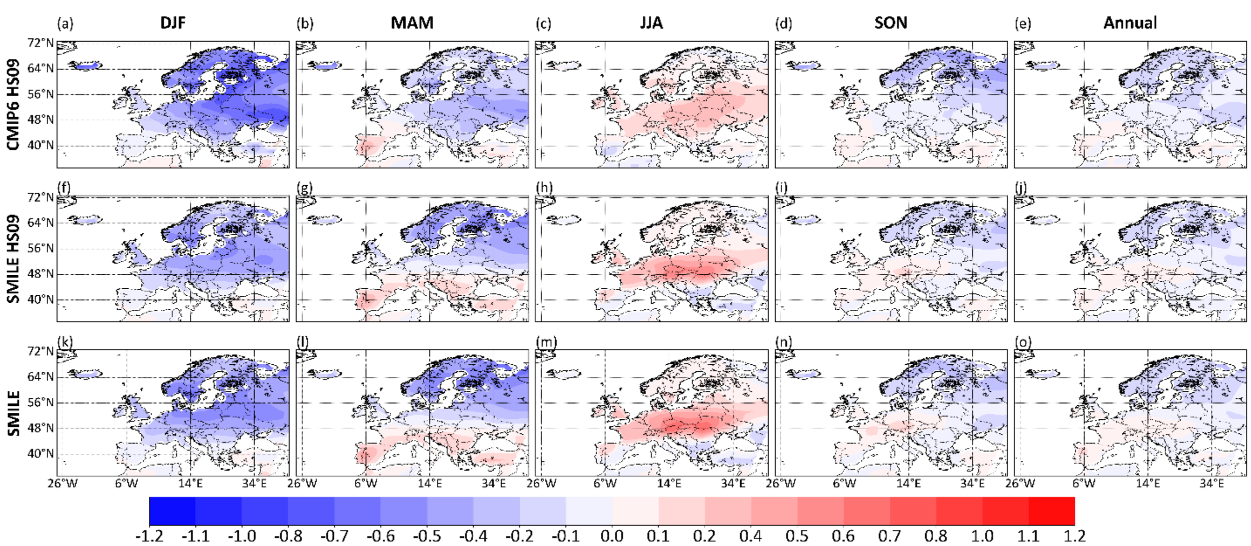

Figure 4: Annual and seasonal differences between the future and historical ICV estimates using the HS09 method for CMIP6 (first row) and SMILE (second row) and using the SMILE TYP method for the third row. Unit: [°C]

Characterization of observational uncertainty

In L21, L323-324 and L350, the claim is made that observational-based estimate are free from modeling uncertainty. Besides the more philosophical point that even statistically-interpolated station data is subject to a "model" and modeling assumptions, this statement is particularly incorrect when applied to the 20CR. Reanalyses are, in effect, based on climate or weather forecast models – they're nudged by the observations through the assimilation scheme, but under the hood they fill the gaps in much the same way a climate model would. This is especially the case for the 20CR, which only assimilates surface pressure (and uses SSTs / sea ice as boundary conditions for their atmospheric model); temperature in 20CR is therefore a purely modeled output.

I'll add that Berkely Earth and GISTEMP were both created for the express purpose of estimating global mean surface temperatures rather than provide a perfect estimate of local temperature. This doesn't disqualify them from their use in this study, but this should be kept in mind. (Note also that GISTEMP does publish an uncertainty ensemble - https://data.giss.nasa.gov/gistemp/uncertainty/ - that could be quite helpful here).

Reply: We appreciate this correction. The sentences stating "observational-based estimate are free from modeling uncertainty" have been completely removed. We have also acknowledged the fact that the observational datasets are aimed mainly to provide a global mean temperature estimate in the discussion: "It is worth noting that the observational datasets used in this study are primarily developed to estimate global surface temperature rather than to perfectly capture local temperature (Rohde and Hausfather, 2020; Lenssen et al., 2024)."

Line-by-line points

L150: I assume the section title should be different here.

Reply: Thank you for pointing out this. Yes, it should be "Results" section and is already corrected.

L192-193: I'm not sure what is meant by "focusing on the methods"

Reply: Thank you for this comment. Based on the updated results, we have reformulated all the interpretations and these words have been changed.

Figures 4-9: could you consider putting all of the dots for a given model in a single x position or connect them with a line? Or maybe just change the colors to be easier to differentiate? I'm having a hard time keeping the different dots and their meaning straight.

Reply: Thank you for this comment. Based on the suggested revisions, the display of Figs. 5-9 has become clearer and not crowded. Therefore, we decided not to add the lines for Figs. 5-9 (neither shaded level as suggested by the first reviewer). However, we have considered this suggestion for the case of Fig. 10, we have added line to connect the same seasons. We have attached the display of the Figures as supplement at the end of this document.

Figures 4-9: Could you also list the SMILE results as dots, like you do with the other CMIP6 models? Either next to the relevant MPI model, or visually separated at the end – but I don't think they're conceptually different enough (see above) to require presenting them like this, and it would help declutter the lines a bit.

Reply: Thank you for the suggestion. We have plotted SMILE as dots.

L282-283: I don't think it's necessarily the case that we would strictly expect ICV calculated over long controls to be larger than ICV estimates based on shorter, current data (since std. deviations show the average spread in a distribution, one would expect better estimates of ICV in a longer time series, but not necessarily larger). I may be misunderstanding – are you instead suggesting that the PI control ICV is likely to be higher because we expect current ICV (which is affected by anthropogenic climate change) is decreasing?

Reply: Thank you for this comment.

The old sentence was from L282-283: "Pidata estimates are based on long time series, so they are robust in this sense, and we expected Pidata results to yield larger ICV estimates than HS09 based on CMIP6 data. The results confirm our expectations in most cases."

The sentences have been changed based on the new results. However, based on the new findings, Pidata 500yrs-STD estimates are higher than those of Pidata 30yrs-STD estimates. This case was not only found in Pidata but in SMILE as well and we have discussed it: "In addition, it is worth mentioning that the ENSMEAN of 500yrs-STD approach systematically produces larger ICV estimates than that of 30yrs-STD method (Fig. 2, 5-9). A similar pattern emerges from the comparison of methodologies applied to the SMILE, with the SMILE TYP approach consistently producing higher ICV estimates than the HS09 method across all seasons and regions (Fig. 2, 5-9). These results are consistent for the spatial and regional means analyses and suggest that longer records tend to increase the magnitude of ICV. This behaviour likely indicates that longer records tend to amplify the estimated ICV, and is likely attributable to the ability of longer records to capture slow modes of ICV that are underrepresented in shorter records."

L305: Since you're looking at seasonal / annual averages, it's not necessarily clear that the day-to-day synoptic variability would have an impact.

Reply: Thank you for pointing out this. Yes, we have corrected the sentence as: "These papers mentioned a common explanation on the fact that it is the future warming which is the main reason for the decrease of temperature variance over many extratropical regions. They emphasized related mechanisms such as polar amplification and reduced equator-to-pole temperature gradients (Screen 2014; Thompson et al. 2015; Olonscheck et al. 2021). Since polar regions warm faster than midlatitudes, the large north-south thermal contrasts is reduced. Moreover, the reduction of sea-ice and snow cover causes stronger warming, which narrows temperature range and reduces variability (LaJoie and DelSole 2016; Olonscheck et al., 2021). These processes weaken temperature variability over the extratropical regions. Additionally, LaJoie and DelSole (2016) found that the anthropogenic forcing leads to statistically significant decline in the future temperature variance over these regions."

L356 – I assume you mean the 20CR data here instead of CHIRPS.

Reply: Thank you for pointing out this. Yes, it is already corrected.

Supplement:

This is the display of Figs.5-9 looks like:

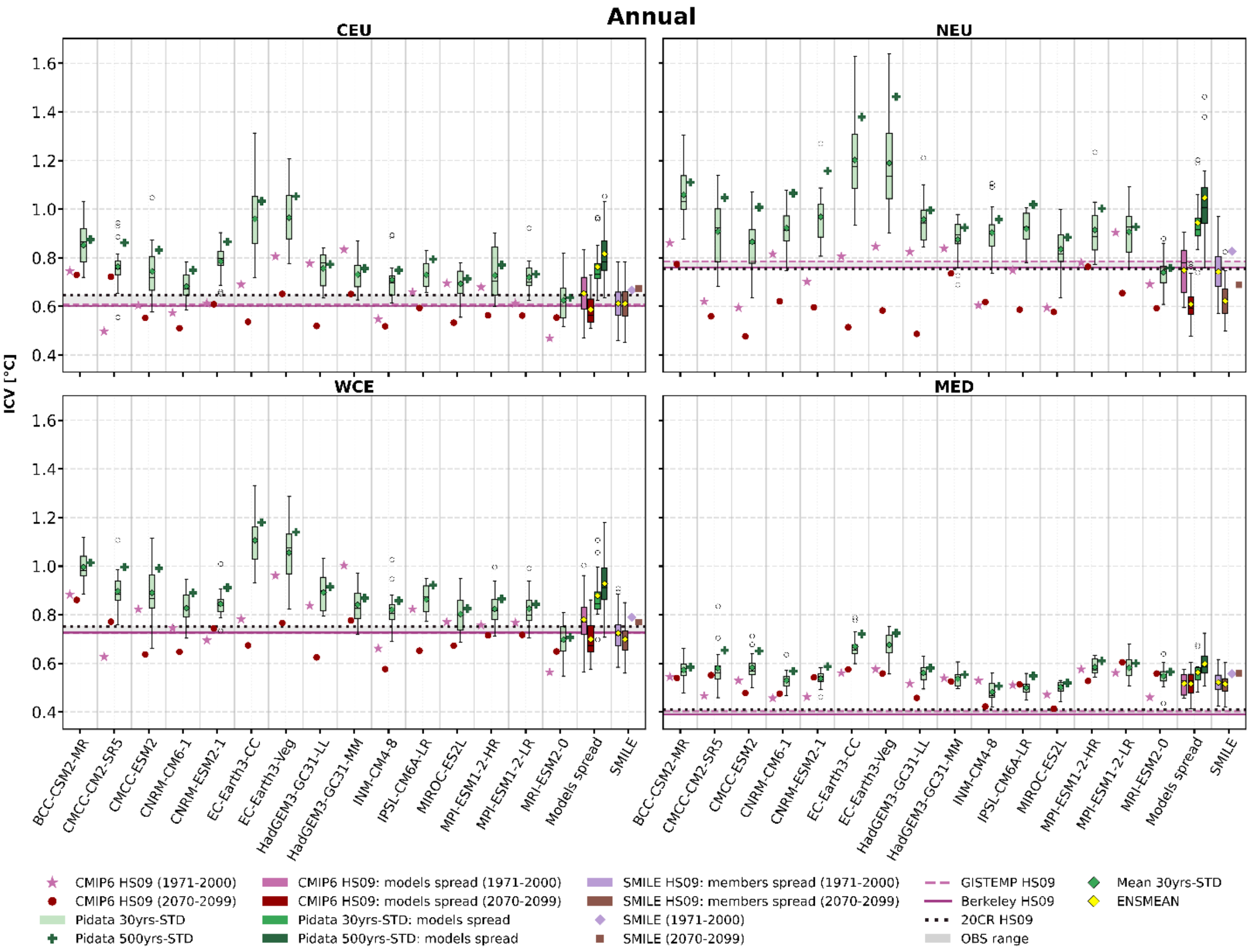

Figure 5: Annual ICV estimates based on the HS09 method applied for CMIP6, SMILE and the observational data, while the STD method is applied in the case of Pidata (represented with different green colors). “Pidata 30yrs-STD” refers to the spread of 16 STD values, each of which was obtained by the application of the STD method for 30 years timeseries. “Pidata 30yrs-STD: models spread” refers to the spread of the 15 CMIP models, in which each model is represented by the mean of the 16 STD values. “Pidata 500yrs-STD” is a single value from the application of STD for 500 years timeseries. “Pidata 500yrs-STD: models spread” refers to the spread of the 15 models based on “Pidata 500yrs-STD”. “OBS range” represents the difference between the maximum and minimum values among the three observational datasets. “Mean 30yrs-STD” refers to the mean of the “Pidata 30yrs-STD”. “ENSMEAN” refers to the multi-model mean as described in the methodology section.

And this is the display of Fig. 10 based on the reviewers’ suggestion:

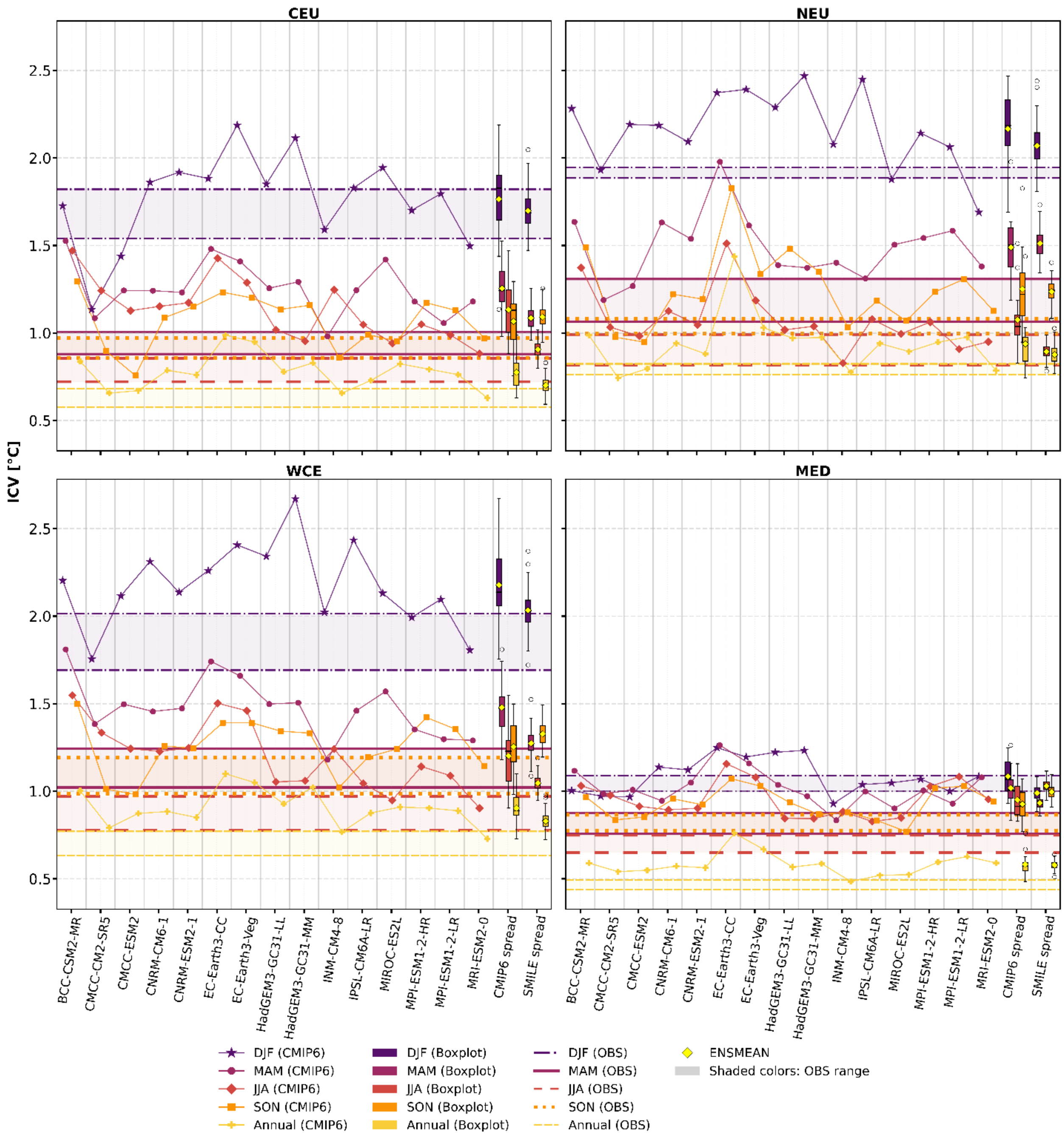

Figure 10: Seasonal and annual ICV estimates based on the REG method during the period 1901-2010 for each of the CMIP6 models, the three observational datasets, and SMILE MPI-ESM. CMIP6 spread gives the spread of the 15 CMIP6 models. SMILE spread gives the spread of the 100 members. The horizontal lines pertain to the observational data, in which the shaded colors represent the difference between the maximum and minimum values among the three observational datasets. OBS refers to the observational data.