ICV paper reviewers' comments

Referee 1:

General Comments

The study analyses Internal Climate Variability (ICV) over the European domain, both on a gridded basis and aggregated over four European regions, for seasonal and annual surface air temperature. It uses climate model projections for different time periods in the historical and SSP5-8.5 scenario, as well as piControl simulations from 16 climate models, a SMILE large ensemble from a single model, and three reanalysis/observation-based datasets. The results from these different approaches to estimate ICV over Europe are then described and compared, with a focus on inter-method, inter-season spatial and inter-model differences. The analysis has potential value for the scientific community, for example by providing quantitative information on current and future ICV over Europe and by evaluating models in terms of their ability to represent it. It may also contribute by comparing different methods for estimating ICV. However, before the manuscript can be published I recommend several major revisions:

In my view, the paper would benefit from a clearer structure based on explicitly defined research questions. Based on my understanding, these could include, for example:

- What are realistic values of ICV over European regions today and in the future?
- How large are the uncertainties in estimating and projecting ICV, where do they come from and how could they be reduced?
- What is the spatial distribution of ICV over Europe, and how does it change from present to future conditions?
- How well do model-based ICV estimates and their spatial patterns reproduce observations, and which models perform best?

At present, the manuscript to me seems to lack a clearly defined set of objectives, which makes it difficult to understand the overarching goal. Clearer research questions would in my view also improve the focus of the results, conclusions, abstract, and introduction parts. In my view, the paper could also benefit from clearly defined metrics for model benchmarking and method comparison, right now the results and conclusions from the analysis appear very imprecise to me.

In my opinion, the text would benefit from more precise formulations and fewer filler expressions (examples below).

The abstract and conclusion appear to include statements that are either obvious or not clearly supported by the analysis (see comments below).

In my view, the methodology is not clearly presented, and as I interpret it, the setup hardens comparability between the ICV estimates from the different methods, as different ICV estimation approaches are applied to different datasets and time periods. The study also seems to lack a discussion of the limitations of the different ICV estimation methods and does not sufficiently address how much of the differences may arise from methodological choices.

The results section appears to largely restates what can already be seen in the figures without much interpretation.

Thank you for all the comments. We address all of them below with specific answers.

Regarding the “benchmarking”, we have defined it 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."

The goal of this present study is not to give a realistic value or range of ICV estimates, nor to assess the performance of the models against observations. Rather, it is to compare different methods for extracting ICV estimates and to be further used for model benchmarking. We have modified the Introduction to explain better the goals of the study.

We have also performed more calculations and analysis, to enable better intercomparison of the results for individual methods. Here are the 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.

Abstract

In my view, the abstract would benefit from beeing clearer on the motivation of the analysis conducted in the study, what do you want to reach with the main innovation of this study, which is the calculation of ICV over europe with the different methods, and what can the reader learn from it or how can he or she apply it? Answering those questions should furthermore focus on the innovation of the study, which is the calculation and comparison of ICV estimates over europe with the different methods and not on ICV quantification in general. Also, in my view, the abstract would benefit from including some of the most relevant quantitative findings, for example the range of quantitative estimates of ICV over Europe and selected results from a model evaluation exercise. This could include the number of models analysed, as well as which models exhibit the highest, lowest, and most comparable ICV relative to the observational estimates (If I understand correctly that this is what you mean by “model benchmarking” in this study). The same, in my opinion, holds for the conclusions part.Some of the conclusions in the abstract could be formulated more cautiously and robustly. The statement “Across all methods and datasets, we found that ICV estimates decrease during the seasonal course from winter to autumn and spatially from north-eastern to south-western Europe. By comparing the results of the historical and scenario simulations of the large ensemble and selected CMIP6 models, we conclude that European ICV generally decreases under anthropogenically forced climate change." is too general. There are several examples where ICV increases under forcing (e.g. the CMIP6 multimodel mean for JJA over Ukraine). Likewise, the decrease from winter to autumn is not consistently observed across all regions and datasets (e.g. Eastern Europe in the CMIP6 multimodel mean for 1971–2000 from JJA to SON), even though this tendency is present in many cases.

At the same time, I would suggest removing statements from the abstract that are not sufficiently explained or interpreted in the manuscript. For example, I don't see where the statement "Moreover, our study suggests that applying ICV estimates as a benchmark for assessing regional climate simulations over Europe should be approached with caution." is backed, explained or interpreted in the remainder of the study.

Similarly, some statements appear to reflect rather general methodological observations instead of conclusions emerging from the presented analysis. For example, the statement "The estimate based on the pre-industrial control simulations offers an advantage since the simulations are not influenced by external forcings and their ensemble mean estimate encompasses the range of the other methods." mainly describes a general property of pre-industrial control simulations rather than a finding of the study itself. In addition, I could not find a clear interpretation or explanation of the statement that their ensemble mean estimate "encompasses" the range of the other methods, beyond a repetition in the conclusions.

Likewise, the statements "When the focus is on future climate simulations, estimates from scenario simulations should be used, as they already account for the influence of anthropogenic forcings on ICV. Regarding ICV estimates from observational data, their advantage lies in accounting for true climate history, free of modelling uncertainty." appear to consist of fairly general and expected methodological observations rather than conclusions directly supported by the analyses conducted in the study, and on top the statement that the observational data is "free of modelling uncertainty" is not supported by the results in the study, where three different observational datasets yield three different estimates for ICV, as those datasets are differently curated, e.g. by using a model constrained by observations as done in the case of Twentieth Century Reanalysis dataset or interpolating between observations.

Also, you are introducing the word "forcing attribution" in the abstract but do not mention it again in the methodology and the statement "Historical simulations also account for historical climate change and yield ICV estimates comparable to those from observational data." is, in my view, way to general and also not really backed by the results of the study, when interpreting the word "comparable" as "roughly equal" as there are many examples where the ICV of the historical simulations fall way outside the observed ICV range (e.g. for MPI-ESM1-2-LR for DJF in NEU, but there are several other examples).

Reply regarding the Abstract: Thank you so much for these comments. We have significantly modified the Abstract to be more specific, clear and to reflect on the new findings. Here is the revised abstract:

Internal climate variability (ICV) represents the intrinsic fluctuations within the climate system. Understanding how ICV varies across regions, seasons, and climatic conditions is crucial for detecting externally forced climate change and for assessing future climate risks. It can also provide a useful benchmark for assessing climate model performance and the emergence of anthropogenically forced climate change. This study aims to compare the estimates of ICV obtained using different methods and types of data. . We focus on seasonal mean near-surface air temperature over Europe. We utilize data from 15 CMIP6 models, a 100-member MPI-ESM large ensemble, and three observational datasets. We compare ICV estimates based on pre-industrial control simulations (Pidata), the Hawkins and Sutton (2009) residual-based method (HS09), regression-based forcing removal, and a single-model large-ensemble framework. The results reveal several key findings. The spatial pattern in ICV estimates generally decreases from northeastern to southwestern Europe on both annual and seasonal timescales, except during JJA, when the highest estimates shift southward. The highest (lowest) ICV estimates are observed in DJF (at the annual timescale). In most cases, Pidata generally yields the highest ICV estimates across seasonal and annual timescales, except during JJA. Based on HS09 method, a stronger agreement is found between the the model-based estimates and the observations' on DJF and annual scales. However, the agreement is weaker in regression-based method method, in which the model-based estimates exceed those of observations in most cases. The projected changes in ICV show a tendency of reduced (enhanced) ICV in DJF (JJA). Although, a north-south contrast emerges at other timescales, with declining (increasing) ICV in northern (southern) Europe, the overall spatial pattern indicates a reduction in projected ICV estimates across large parts of Europe, except during JJA.

Introduction

The introduction discusses uncertainties in climate projections, defines ICV and its origins, and presents several methods previously used to quantify ICV and publications which use them. However, I think the introduction would benefit from the following improvements:

(1) The current discussion places comparatively strong emphasis on uncertainties in general (including model and scenario uncertainty), while the relevance of ICV quantification itself receives less attention. I would suggest reducing the focus on the broader uncertainty discussion, as this is nothing the paper focuses on, and instead elaborating more on why quantifying ICV is important, as currently only one concrete example is provided.

(2) In my view, the manuscript would also benefit from a clearer explanation of how the comparison of different methodologies and datasets contributes to a better understanding of ICV

and to model benchmarking. At present, the introduction mainly states that several methods are combined to “compare and assess ICV magnitude from diverse perspectives” and to “provide a useful benchmark” for evaluating model performance, but it remains somewhat unclear to me how exactly these comparisons contribute to these goals.

In addition, I think it would be valuable to formulate more explicit research questions in the introduction that can then be addressed throughout the paper using the different methods and datasets.

Reply regarding “Introduction”:

We have restructured the Introduction. We start with explanation of ICV itself, put it into context of climate modelling (the uncertainties), and provide an overview of methods previously used for ICV evaluation. We also better explain the idea of the benchmarking the climate model outputs with the ICV estimate:

“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. "

We also further emphasize the importance of ICV changes under anthropogenic forcings. In the last paragraph, we then state specific aims of the study: "Despite the wide range of approaches, a systematic comparison of ICV estimates across methods and datasets, including observational constraints, remains lacking—particularly at the regional scale over Europe. Given that each of the above mentioned methods and datasets play a valuable role in assessing ICV, this study combines, for the first time, several approaches and datasets to compare and assess ICV magnitude from diverse perspectives. We focus on the annual and seasonal means of near-surface air temperature. Specifically, the study aims to quantify historical and future ICV levels across Europe, to evaluate the magnitude of uncertainty in ICV estimation using different approaches, and to characterize the spatial distribution of ICV over Europe and its evolution under future conditions. "

Regarding the "forcing attribution", we decided to basically avoid this term, as it may be a bit misleading. Now we use the "REG method" abbreviation as explained above, and the "attribution" is mentioned only once in the Introduction with appropriate references to literature: "Another approach to quantifying ICV magnitude is to look into the observed climate variability and to attempt its attribution to external and internal forcings (e.g., Foster and Rahmstorf, 2011; Mikšovský et al. 2016). If we are able to estimate the influence of external forcings, and subtract the forced signal from the observed time series, the resulting residuals represent the estimate of ICV. "

Methods

In each paragraph describing the methods, or alternatively in a short overview at the beginning of the section, the order of the calculations could be stated more clearly. In particular, it would help to clarify in which order detrending, regridding, temporal aggregation, spatial aggregation, and standard deviation calculations are applied. For the HS09 method specifically, it is currently unclear whether the different time intervals are separated first and then detrended, or whether detrending is performed before the separation into periods.

Reply: We have clarified step by step how the calculation in each method in the methodology section. We have divided the section into two parts – we first introduce the data, and then elaborate on each of the methods with detailed explanation.

If the periods are first separated and the 4th-order polynomial detrending is then applied to individual 30-year periods, the polynomial fit may smooth out part of the low-frequency variability together with the forced signal. This should be checked and discussed

Reply: This is true: Long-term variability will be suppressed by the application of the given polynomial fit, even if its originates from internal variability (e.g., AMV/AMO variability is often considered to have a dominant near-periodic component of about 70y, so within a 30-year window, this may manifest as a trend-like structure, and filtering by a polynomial will remove most of it). The application of 4th-oder polynomial is based on the HS09 method itself. The method is described as follows: “First, 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." This given algorithm is restricted to high-to-mid frequencies, as long-term variations become difficult to distinguish from long-term components of external origin.

It might be good to clarify whether area weighting is applied during the spatial aggregation, i.e. whether the differing spatial areas represented by individual grid cells are taken into account.

Reply: Thank you for pointing this out. We performed spatial aggregation with “cdo fldmean” and it accounts for weighting (Schulzweida, 2022).

It would also be useful to discuss to what extent the ICV estimates derived from the four methods are directly comparable in terms of which processes and forcings are included or excluded. For example, in the observational analysis, volcanic aerosols and solar variability appear to be implicitly removed together with the forced trend. In contrast, in the SMILEs, these forcings would also be removed because they are identical across ensemble members and therefore cancelled out through subtraction of the ensemble mean. However, in the other methods, volcanic and solar forcing effects may still contribute to the estimated variability because the applied trend removal does not explicitly account for them. This may potentially also contribute to the comparatively larger variability estimates obtained from the PiControl simulations. In my opinion, the manuscript would benefit from a clearer discussion of which effects are included and excluded in the ICV quantification of each method and how this affects the comparability between methods.

Reply: For better comparing ICV between different datasets and methods, we have applied linear regression (REG method) to CMIP6 and SMILE, in the same way as we did for observations. We also applied HS09 not only to CMIP6 and SMILE but to observations as well. These enable us to compare the different data while using the same method. Moreover, we have added discussion of the differences between all the methods into the last section of the paper: "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."

And in another paragraph, we added: "On the other hand, the SMILE TYP method explicitly removes the external forcings by subtracting the ensemble mean from each ensemble member. This procedure more effectively isolates the ICV. As the SMILE TYP method generally yields higher ICV estimates than the SMILE HS09-based approach, we conclude that in this case the

HS09 erroneously ascribes a part of internally caused variability to the externally forced component estimated by the 4th order polynomial."

Also: "Moreover, it is noticed that the ICV estimates based on REG method are higher than those of HS09-based. In addition to the influence of the longer time series used in the REG method, the higher ICV estimates obtained with this approach may also be related to the way of removal of the influence of the external forcings, as discussed above. 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."

The application of linear trend removal to the piControl simulations should also be discussed and checked, as this implicitly assumes that model drift is approximately linear. It is unclear whether this assumption is generally valid. If model drift contains nonlinear components, residual drift effects may remain in the ICV estimates, which could potentially contribute to the systematically larger ICV values obtained from the piControl simulations.

Reply: We have also investigated temporal trends and detected no notable temporal trends in any of the Picontrol time series. However, removing nonlinear trends is out of the scope of this present study so far. We wanted to estimate the ICV from Pidata based on the previous method used by Olonscheck and Notz (2017). Previous studies have, as far as our knowledge, used mostly linear trend for the drift removal. So we align with this approach and remove the potential drift in this way.

Similarly, the use of a linear assumption between forcing and regional temperature for the observational datasets may also be problematic. While the relationship between global forcing and global mean temperature is approximately linear, this does not necessarily hold at the regional or grid-point scale. Regional temperature responses may contain nonlinearities and path dependencies associated with circulation changes or feedback processes such as albedo changes from changing snow cover. This assumption should therefore be tested and discussed more explicitly.

Reply: Indeed, the relationship between the forcings and local climate variables should generally be considered nonlinear. However, for such nonlinearity to manifest in a detectable (and statistically relevant) manner, a substantial amount of data is typically required, sufficiently capturing the nonlinear components of the relevant relationships. In our case, the target period is too short for reliable application of nonlinear regression methods (i.e., to properly estimate the higher number of free parameters of the regression model), thus a linearized version of the model has been employed as a more robust alternative.

It is also unclear to me whether the subset of models used for the ICV evaluation is representative of the broader CMIP6 model spread, for example with respect to TCRE or other measures of climate sensitivity and model uncertainty. Selecting models primarily based on the availability of piControl simulations over a given period appears somewhat arbitrary and could potentially introduce sampling biases into the ICV estimate from the ensemble mean (e.g. if only hot models are sampled).

Reply: Thank you for this comment. The subset choice is driven primarily by data availability, but it also represents relatively wide range of model families and ECS values. We have added that in the method section: "The choice of the 15 selected CMIP6 models is based on the availability of pre-industrial control simulations, each providing a continuous 500-year run from 1850 to 2349. Moreover, even though 15 models were chosen, the subset represents half of CMIP6 model families (7 of the 14 CMIP6 model families) (Brunner et al., 2020), with equilibrium climate sensitivities ranging 1.8°C to 5.6°C (Meehl et al., 2020).

Using different time periods across the datasets, could further reduce the comparability of the results. For example, the observational datasets estimate variability over the 1901–2010 period, whereas the historical simulations and SMILEs are analysed over 1971–2000. Since the external forcing conditions differ substantially between these periods, differences in the estimated ICV are expected, which makes direct comparison and potential "model benchmarking" based on the provided data more difficult.

Reply: We appreciate this comment. As already described above, we have performed additional analyses and applied the same time periods for SMILE, CMIP6 and observations. For the case of HS09 method, we applied 30 years period (1971-2000). And for linear regression method, we applied 110 years (1901–2010). We have included this explanation in the Methodology section.

In addition, the large differences in the lengths of the analysed periods should be discussed more explicitly, as they may introduce substantial sampling uncertainty. Calculating a standard deviation from 30 years of data is considerably more uncertain than estimating it from several hundred years, especially if the ICV is correlated through time, an effect that could explain the differences between the ICV estimates based on PiContorol and the other datasets. One could also test this by additionally calculating the ICV from selected periods within PiControl that match in length with the other data.

Reply: We have added the calculation of 30 years period for the case of Pidata and have discussed how different length could contribute to uncertainty in the results. Based on the results, we mention: "In addition, the ENSMEAN of 500yrs-STD approach systematically yields larger ICV estimates than the ENSMEAN derived from 30yrs-STD method." And "These suggest … that longer records tend to amplify the estimated ICV, likely reflecting an improved sampling of low-frequency variability."

To further improve the comparability between methods, it may also be useful to apply the different approaches more consistently across datasets. For example, the HS09 method could additionally be applied to the piControl simulations and observational datasets, while the forcing-attribution approach could also be applied to the piControl and historical simulations. Similarly, applying both the HS09 and forcing-attribution methods to the SMILE simulations could help disentangle how much of the differences reported in the study arise from the methodological choices themselves versus differences in the underlying datasets/models. This would substantially improve the interpretability of the different ICV estimates and enable more angles of comparison.

Reply: We have addressed all your comments regarding the methodology. Here are our overall comments and description of the changes regarding the methodology:

We have restructured the section about Data and methods. We first introduce the data we are using and then describe the methods. We also enhanced the description of the methods, specifying more details about the calculations.

Most importantly, we have performed more calculations, to make the results more comparable, as suggested. As mentioned earlier, we have applied the HS09 and linear regression methods to CMIP6, SMILE simulations and observations. However, applying HS09 to Pidata would be inappropriate since the preindustrial simulations contain only variability stemming from internal fluctuations.

Additionally, as mentioned earlier for the case of Pidata simulations, we adopted the recommendation, and we added the ICV estimate based on 30 years' time period. So, the time periods now align better with that of HS09 and SMILE typical method based on Lehner et al., (2020) results.

Results (4)

This section mainly describes patterns, quantifications, and observations that can be directly inferred from the figures. However, in my view, it would substantially benefit from more interpretation and discussion of the underlying physical mechanisms behind the reported results, some of that also follows in the conclusion section, but I think it would make sense to add it here already.

Reply: We have largely revised the description of the results. Now the text is more focused on describing general patterns and pointing out interesting details. We have also added some sentences discussing some of the results, or offering hypotheses about the reasons for the revealed patterns. For example, " the difference among the observational datasets is smaller than that among the model-based datasets across all regions and timescales (Figs. 5-9). In addition, the inter-model spread of CMIP6 generally exceeds the spread among SMILE members in most cases. These findings suggest that uncertainties associated with observational datasets are relatively minor compared with those arising from climate model structural differences, hile the uncertainties stemming from structural variations across climate models are larger than those resulting from multiple realizations of a single model." Still, most of the discussion is done in the Section 5.

The subsection titles need to be revised. This is clearly part of the Results section rather than "Methods and data". In addition, the title of Section 4.2, "Spatial distribution of ICV estimates", does not fully reflect the actual content of the section. A title such as "Comparison of spatilly aggregated ICV estimates across methods" or similar would better capture the focus of the analysis.

Reply: There were errors in some of the subtitles, we are sorry for that, we have corrected them as "4.Results" and "4.2 Regional mean analysis of ICV estimates"

The methodological concerns raised in the Methods section would also directly affect the interpretation of the results. In particular, the different methodologies should be hard to compare, as they include and exclude different sources of variability and forcing effects, are applied over different periods, and use substantially different temporal sample sizes. In my opinion, these limitations should be discussed more explicitly when comparing the resulting ICV estimates, or better removed from the analysis by revising the methodology and e.g. matching timeperiods.

Reply: We have performed additional calculations and matched the time period as described above. We have also added more discussion of similarities and differences between the results of individual methods, which have been enabled by the newly performed calculations.

More generally, several formulations throughout the Results section appear too broad or imprecise. An example among others is the statement "The annual analysis yields smaller ICV estimates than the seasonal analyses, while DJF shows the largest ICV magnitudes across all seasons. Pidata provides the highest ICV estimates among all the approaches and timescales, except during JJA. GISTEMP shows lower ICV estimates than the other results, while Berkeley and 20CR display comparable ICV estimates in most cases." is only true in a broad overall sense, but there are numerous exceptions to each of these claims. For example, there appear to be regions and methods where the annual analysis does not yield the lowest ICV values (e.g. northeastern Europe in the historical SMILE simulations). Likewise, DJF often exhibits the highest variability, but not consistently across all regions and datasets (e.g. Iberia in the SMILE scenario simulations for 2070–2099). Similarly, piControl simulations often show the highest ICV values, but there are exceptions (e.g. Iberia in MAM compared to the SMILE historical simulations). The statement regarding GISTEMP also appears too broad, as there are regions where GISTEMP produces comparable or even larger estimates than other datasets (e.g. DJF over northern Italy). Finally, the statement that Berkeley Earth and 20CR are "comparable" is somewhat unclear to me in its current form. Presumably, the intention is to state that they show relatively similar spatial and seasonal patterns and magnitudes across many regions.

Reply: These sentences have completely changed based on the new results and the revision of the text. And the description of the new results have been adjusted and clarified more. For example, we have specified the location ("northwestern Europe") in which Pidata displays the highest estimates. For example: "Pidata STD methods (Fig. 2) generally produce the largest ICV estimates across the different approaches and timescales over the northwestern Europe, except during JJA."

Also: "The smallest ICV estimates are depicted during the annual scale over the MED region, while the largest estimates are found in DJF over the northwestern Europe." These are only few examples, the revised version of the manuscript has changed markedly from the originally submitted text.

The statement "In general, despite the differences between the methods, there is good agreement between the modeled datasets and the observational ones. However, it is noticed that most of the modeled datasets display noticeably higher ICV estimates than the observational ones during JJA over all the regions, except the NEU." also appears difficult to justify in its current form and I don't think it is well defined what good means here. In many cases, the differences between models and observations are quite substantial (e.g. differences exceeding 50% in MAM over WCE for EC-Earth3-CC). Therefore, the conclusion of "good agreement" should, in my view, either be supported more quantitatively or formulated more cautiously.

Reply: Thank you for this comment. These sentences have been completely changed based on the new findings.

Conclusions:

As in other sections, in my view, the conclusion contains several imprecisely formulated statements. An example is "Our results imply that the estimate based on the ensemble mean value of preindustrial control simulations (denoted here as Pidata ENSMEAN) encompasses the range of the other methods," where it is unclear to me what is meant by "encompasses the range" in terms of metric or interpretation.

There are also statements that to me do not clearly follow as conclusions from the analysis but rather read as general observations about the datasets, for example "So, when the focus is on future climate simulations, the estimates involving scenario simulations should be used, as they already account for the influence of anthropogenic forcings on ICV."

Overall, the conclusion would in my opinion benefit from more precise formulations and a clearer distinction between results derived from the study and more general statements and, as the rest of the paper, from clearly defined research questions and evaluation metrics. Also it should discuss methodological limitations of the study which could influence to the observed results.

Reply: Thank you for pointing this out. The conclusion has been significantly revised based on the updated results and all comments from the reviewers. Moreover, the novelty of the study has been stressed out and the limitations of the study have been described. The Section 5 "Discussion and conclusions" is now structured in several paragraphs. First, we describe the spatial and seasonal patterns of ICV across Europe and compare them with previous studies, and also describe potential explanations of revealed patterns; second, we evaluate the strengths, limitations, and assumptions of the different ICV estimation methods; third, we compare the resulting ICV magnitudes across methods, models, and observations; fourth, we assess historical versus future changes in ICV and discuss the underlying physical mechanisms; fifth, we examine differences among observational datasets and between CMIP6 and SMILE simulations. In the final part, we summarize the main findings, listing bullet points with the most important conclusions of the study:

- In most cases, Pidata provide the highest ICV estimates (except in JJA), because they combine longer simulation records and absence of external forcings.
- The HS09 method reveals that the ICV estimates derived from CMIP6 and SMILE generally agree well with observational estimates at the annual and DJF timescales across most regions. In contrast, during MAM, JJA, and SON, the model-based estimates tend to exceed those derived from observations. In most regions and seasons, the REG method indicates that model-based estimates exceed those inferred from observations. This suggests that either climate models simulate stronger regional-scale ICV or the observational estimates do not fully capture ICV because they are based on gridded temperature reconstructions rather than perfect observations of local temperature.
- Both historical CMIP6 HS09 and SMILE HS09 reveal better agreement at the annual and DJF timescales than during the other seasons. However, their future counterparts exhibit agreement only at the annual timescale. In REG method, CMIP6 ENSMEAN generally produces higher ICV estimates in most cases than the SMILE ENSMEAN across most regions on the seasonal and annual timescales.
- The inter-model spread of the CMIP6 is generally larger than the spread among the SMILE members, suggesting that structural differences between climate models generate greater uncertainty than multiple realizations from a single climate model.
- The differences in ICV estimates between the observational datasets are typically lower compared to differences between individual CMIP6 models. In this sense, the observational uncertainty is lower than the model uncertainty.
- The projected changes in ICV are strongly season- and region-dependent, with a tendency towards reduced ICV in DJF, but enhanced ICV in JJA. However, during the transitional seasons (MAM and SON) as well as on the annual timescale, a clear north-

south contrast emerges, with decreased future ICV estimates over northern regions and increased future estimates over southern regions compared with historical conditions. The overall spatial pattern indicates that reductions in projected ICV estimates dominate across large parts of the study area at both seasonal and annual timescales, except during JJA.

We conclude with discussion of limitations of our study, and implications for future applications of ICV estimates in climate-model evaluation:

"The better agreement between the model-based and observational data in the HS09 method than in the REG method might imply that HS09 is more robust than REG. However, our study has a limitation due to shorter timeseries in HS09 (30 years) compared in REG (110 years). Moreover, it should be noted that in this study, we only used one SMILE simulation originated from a single climate model. This constitutes another limitation of the study and should be considered when interpreting the comparison between the SMILE and CMIP6 ensembles.

Regarding the potential use of ICV estimates for benchmarking and assessment of climate model outputs, our results emphasize that the ICV estimates differ between seasons, regions and models. Therefore attention should be paid to selection of the ICV estimate method to ensure methodological consistency. If possible, estimating the ICV for the time period and data source of interest seems most logical (e.g., inferring the benchmark for the climate model of interest from that particular model). "

Figures and Tables:

It is not clear to me which method is used to determine whether a grid point is displayed in Figures 2 and 3. In particular, some of the grid points in coastal regions do not appear to align consistently with the land–sea mask shown in the figures.

Reply: Thank you for this comment. We have added this explanation in the manuscript: "For the comparison purposes, we regridded all the results into common grid of 1° x 1° (longitude × latitude) using bilinear interpolation in Climate Data Operators (CDO, Schulzweida, 2022). More specifically, we first computed the ICV estimates as described in the methodology section (3.2 Methods) for each of the datasets at their original grid and then regridded all of them into the common grid. It should be noted that the mask dataset has a relatively coarse spatial resolution (1° × 1°). As a result, the spatial extent of the masked regions may appear inconsistent with the actual coastlines, represented by the thin country-border lines in the figures (Fig.2-4)."

In Figures 4–8, it would be helpful to include the shading levels in the legend to improve readability and interpretation.

Reply: Based on the suggested revision, the display of Figs. 5-9 (previously it was Figs.4-8) has become clearer and not crowded. Therefore, we decided not to add the shaded level for Figs. 5-9 (neither lines as suggested by the second reviewer). However, we have considered this suggestion for the case of Fig. 10, we have added lines to connect the same seasons. We have attached examples of the revised Figures at the end of this document.

In Table 1, it is not always clear to me which model name corresponds to which modelling centre description, and some entries appear incomplete (e.g. MIROC-ES2L). Adding clearer separation between rows and columns, for example by using lines, could improve readability.

Reply: Thank you for this comment. We have adjusted the table.

Code:

The code is not provided in a public repo so I couldn't check it.

Reply: Thank you for this request. So far, the scripts developed for this study have not been publicly archived because they are part of an ongoing research project. However, we have mentioned in the manuscript that they are available from the corresponding author upon request.

Conclusion

Overall, the manuscript addresses a relevant topic and has the potential to contribute useful quantitative insights into internal climate variability over Europe and its estimation using different methods, datasets and models. However, in its current form, the study suffers from a lack of clearly defined objectives, limited methodological comparability across datasets and approaches, and insufficient discussion of key limitations. In addition, several conclusions appear either insufficiently supported or too general/imprecise in their current formulation, and the results section would benefit from more interpretation rather than primarily descriptive summaries. I therefore recommend major revision before the manuscript can be considered further.

Reply: Thank you for all your comments, we find all them very useful. We have significantly revised the whole manuscript based on the reviewers' suggestions as described above. We believe that we have addressed all the comments and improved the quality of the manuscript.

Supplement: Examples of the revised design of the figures, implementing all the newly calculated results.

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

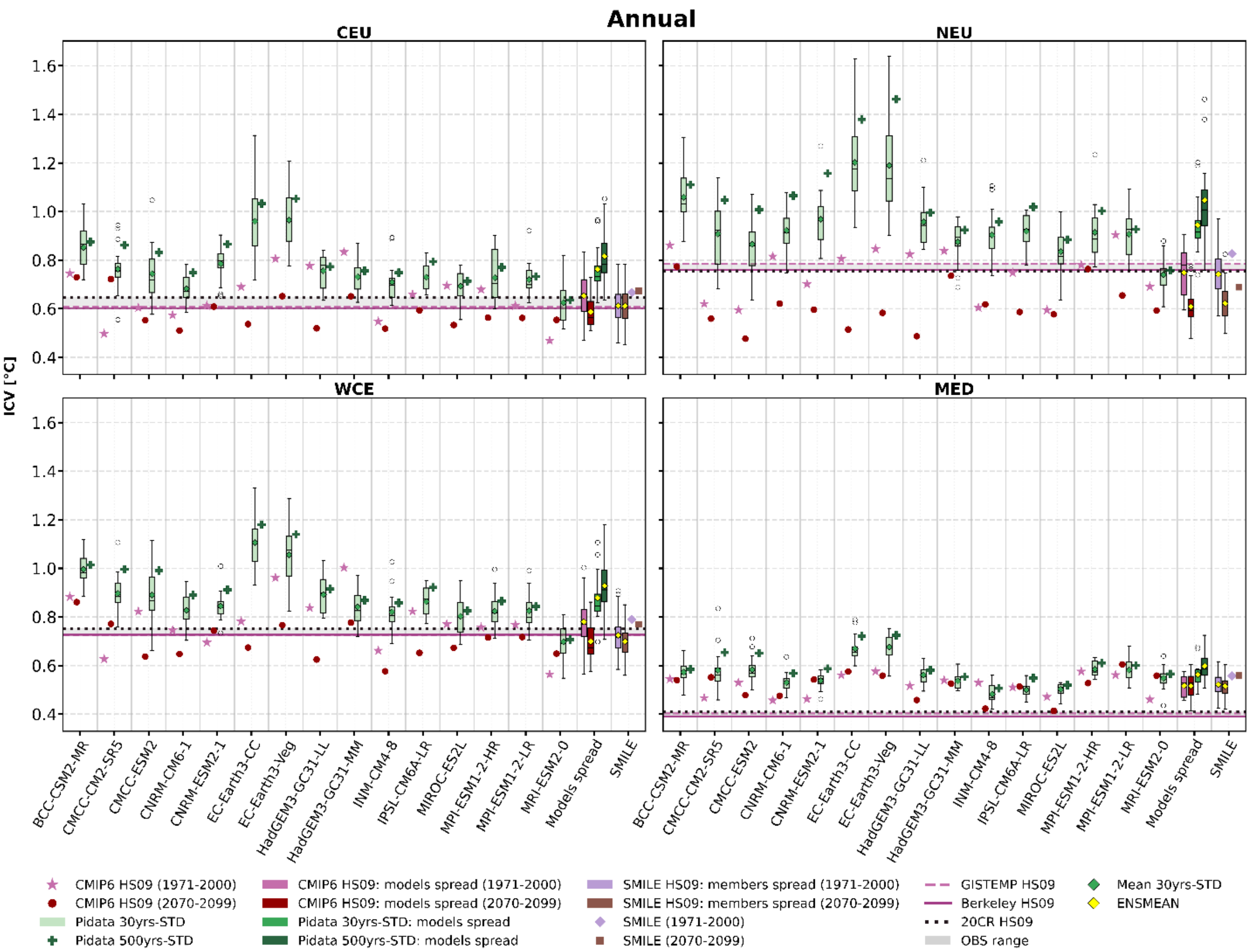

Figure 5: Annual ICV estimates based on the HS09 method applied for CMIP6, SMILE and the observational data, while the STD method is applied in the case of Pidata (represented with different green colors). “Pidata 30yrs-STD” refers to the spread of 16 STD values, each of which was obtained by the application 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:

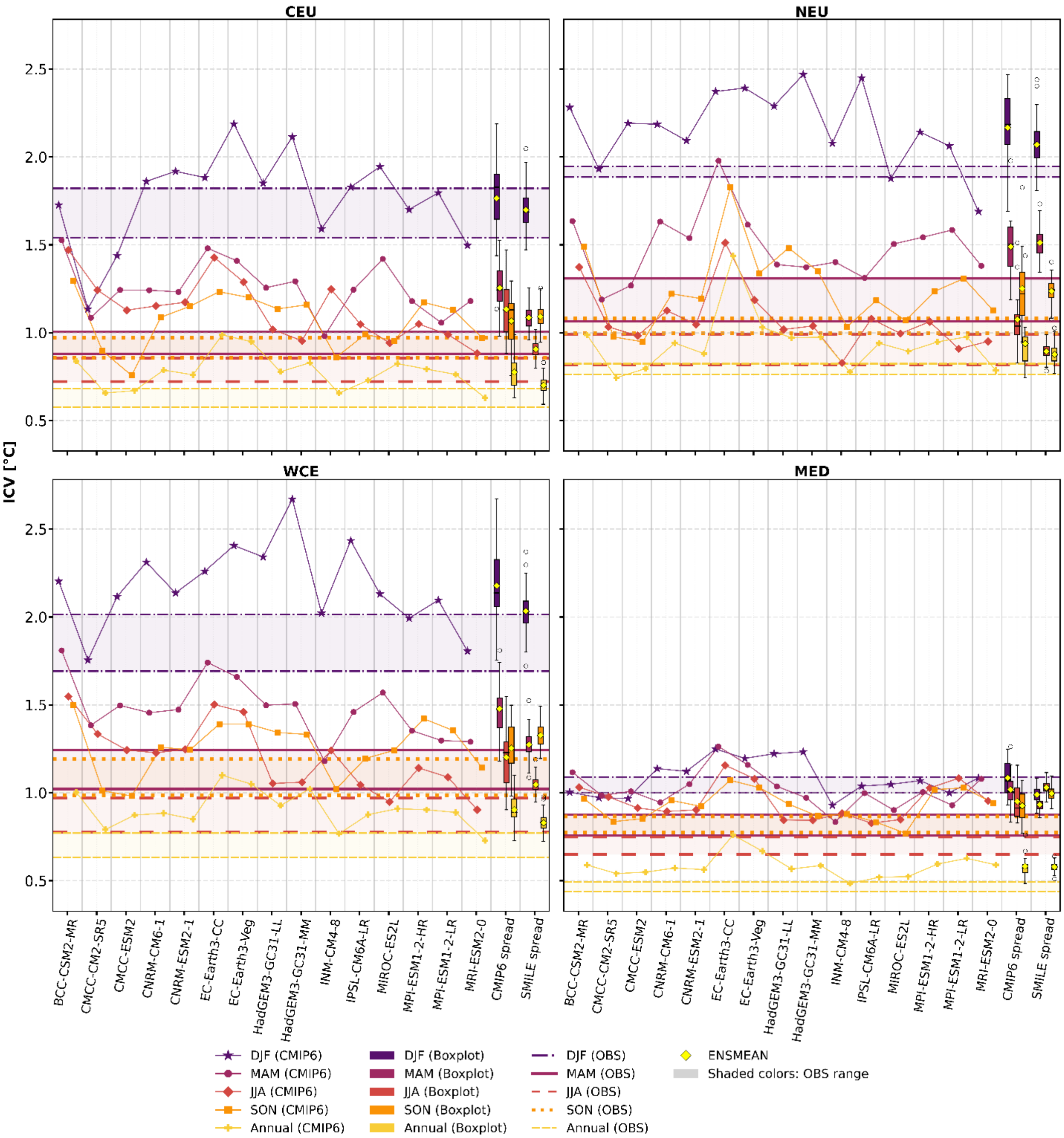

Figure 10: Seasonal and annual ICV estimates based on the REG method during the period 1901-2010 for each of the CMIP6 models, the three observational datasets, and SMILE MPI-ESM. CMIP6 spread gives the spread of the 15 CMIP6 models. SMILE spread gives the spread of the 100 members. The horizontal lines 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.