Response to Reviewer Comment #1

30.04.2020

First we want to thank you for your long and detailed examination of our manuscript.
Some general remarks from our side before we will comment on every paragraph individually:

Both reviewers raised concerns on two major aspects:
1) confusion about our use of the terms internal variability (IV), inter-annual variability (IAV) and inter-member variability IMV). We carefully define our use of the terms and make the separation clearer in the revised manuscript, and thereby take up your valuable comments on our approach to treat IMV as an approximation for IAV; we still find the concept appealing (and they lead to very similar results as can be seen), but we also understand your concerns about it, especially when it comes to the interpretation of results.
2) the lack of a clear line of argumentation. We will stronger focus on the IAV (compared to E-OBS and how it will develop in the future in the 3 SMILEs) throughout the manuscript, and add a chapter on the connections of IAV and IMV at the very end. This controversial topic is thus shifted away from the main line of argumentation.

Another important point, especially regarding your comments on further analysis of the driving GCMs and driving mechanisms: The paper is written from an RCM user and impact modeler's perspective, the background of the first author, rather than an atmospheric sciences/model developers perspective. This was obviously not clear enough in the current version and we now emphasize the relevance for the RCM and impact modeler community in the introduction and discussion of the revised manuscript. The revised paper is stronger focused on this perspective, including a more detailed examination of the influence of the model biases in the mean climate state on differences in inter-annual variability in the comparison against E-OBS. This means that we will extend the analysis from the last paragraph of section 5.1 in more depth, as this is very important information for impact modelers, who usually need to bias adjust RCM data.

I have difficulties in giving a condense summary of the study by von Trentini and colleagues. As I understand the present ms, the authors intend to investigate the effect of internal variability on projected changes in inter-annual variability of key atmospheric variables over Europe by means of 3 high-resolution SMILEs. This would be scientifically interesting and would provide new and important insights into the uncertainty of simulated future climate change signals. However, the concept of inter-annual variability is treated here as a concept of internal variability (often synonymously), thus causing a lot of confusion of concepts and distraction from a clear line of investigation and argumentation. I dont see an advantage in doing this. Why not investigating projected changes in inter-annual variability the same way as projected changes in any other variable, using i) the SMILEs to provide a sound estimate of the associated uncertainty due to internal variability (i.e. sensitivity to initial conditions), and ii) the three models to provide an estimate of the associated uncertainty due to the choice of the climate model? The problem becomes apparent already in the introduction (L34ff) when uncertainty in projected climate change signals due to internal variability is confused with "future changes in uncertainty due to internal variability". The mixture of internal and inter-annual variability sometimes leads to unsound comparison, e.g. L34 and 37 (see below), and even to unsound conclusions, e.g. that an increase in inter-annual variability implies an increase in the uncertainty of climate projections due to internal variability (L338).

Generally, we will more strictly focus on variability on annual timescales and remove all references to uncertainty in climate projections, which appeared to be confusing. Further clarifications of terms are

given in some of the specific comments later.
We took up the suggestion to treat IAV just like any other variable and changed the whole methodology for examining future changes in IAV (former Figures 8 and 9). As the new methodology is strongly connected to your comment on lines 234ff ("Are these changes significant? […]"), it will be presented further down at the respective response. As this is the only part of the manuscript where we actually make use of our assumption that IMV is a good approximation for IAV, all discussion on this point will be shifted to the very end of the manuscript in the revised version. This will make it easier for the reader to follow our line of argumentation. The (obviously) controversial topic can then be discussed in all details without "disturbing" right from the beginning.

Moreover, the presented analyses are sometimes questionable. For instance, the significance test for a linear trend in IMV (Fig.9) is based on time series which have been smoothed by a 20-yr running mean. A running mean can heavily reduce the variance of a time series and thus increase the significance of its linear trend artificially. Anyway, the significance of climate change signals is more meaningfully tested against the variance of an unforced control simulation (constant atmospheric greenhouse gas concentrations). Such a control simulation is not provided for any of the RCMs but would be essential for each to substantiate the results.
We agree that the significance test was not a well suited method to test on significant climate changes in IAV. It was therefore discarded. As unforced control simulations are not available for the corresponding RCM simulations, we cannot use them to make statements about the climate change signals compared to pre-industrial times. However we can detect a robust climate change signal in IAV compared to a historical reference climate.

Further, I miss some important analyses. The results of the investigated RCMs are not compared with their parent driving GCMs (the authors are aware of this, L328). Such a comparison, however, is of high interest and would increase the impact of the study significantly since it allows to assess the error in GCM-based estimates of inter-annual variability. I expect the signal over the British Isles for example to be strongly influenced by the temperature of the ocean, which is prescribed by the GCM in two out of the three RCM ensembles. A RCM-GCM comparison might also provide important information about the influence of the RCM domain size on the projected change signals. The RACMO domain for example is rather small and accordingly I expect a strong influence of the boundary conditions here. Also the imipact of different ensemble sizes is not discussed but of high interest. The ensemble sizes range from 16 to 50 members. Do the results suggest that 16 ensemble members are enough to study internal and/or inter-annual variability in the atmosphere?
We agree that a comparison of the connections and differences of RCM large ensembles and their driving GCM large ensembles is an important research task. However, such an analysis is out of the scope of this paper. Such an investigation would open up a huge new field of analysis that needs to be performed despite just calculating indices for the GCM data. Since the paper primarily addresses the RCM user/impact modeler community, we kept a GCM-RCM analysis explicitly out of the scope for this paper.
Concerning the ensemble size: this is also a very important question that many people ask in the context of SMILEs. We found that the measure we use for the description of variability (standard deviation) is not very sensitive in the range of 16-50 members we have in our data sets. We can add a short section on this. But it is worth mentioning that another paper in this special issue by Sebastian Milinski et al. ("How large does a large ensemble need to be?") is dealing exclusively with the question of ensemble size by means of the 100 member MPI-GE.

Finally, a great advantage of having 3 ensembles at hand is that we can learn a lot about the driving mechanisms of the simulated future changes and their representations in different models. What are the physical driving mechanisms of the changes that agree in sign and what could be the reasons for

the disagreements? The results should be put closer into context e.g. of the studies cited in the introduction (L59-74). Some suggestions of driving mechanisms are already given but should be strengthened by analysing and explaining more details. E.g. L310, arctic amplification and sea ice loss as a driver for decreasing winter temperature variability in Europe is not obvious.
We agree that this is an exciting research question. This point is a bit connected to the GCM-RCM discussion, as we also do not see ourselves doing an in depth analysis of the driving mechanisms by means of data analysis. However, there has been a lot of research done towards exploring the mechanisms driving changes in variability. Thus, we extend the discussion of possible driving mechanisms by more explanations and especially with the references to existing literature.

Because of these major concerns, I suggest to reject the ms in its present form. Nevertheless, because of the great potential that I see in the comparison of 3 GCM/RCM SMILEs I like to encourage the authors to revise/extend their study thoroughly and resubmit a new ms.
With a stronger emphasis on the key points we want to analyze in the paper, clearer distinction between inter-annual variability and uncertainty due to internal variability, and the changes in methodology already (and later on) mentioned we hope to convince you with a more concise and comprehensible manuscript, following a clear line of argumentation.

Other general comments:

It is worth to add to the discussion or conclusions section that the ensemble means of the projected changes can be interpreted as the future changes associated with highest probability (under the considered emission scenario and the individual model constraints) but which specific change would in fact become realized depends on internal variability.
We will pick up this comment and add a sentence or two to clarify this important point.

Please also add that by evaluating simulated inter-annual variability with E-OBS you also assume that this single realization (and period) of nature is not an outlier in terms of inter-annual variability under the prevalent climatic conditions.
We will discuss this. Note that we evaluate whether IAV in E-OBS falls within the range of IAV in any of the ensemble members. Of course the observed variability could still result from a realization of climate that is an outlier but in contrast to most previous studies based on individual simulations we consider our analysis as more robust.

In many paragraphs, the distinction between historical conditions and projected future changes is not clear. E.g. L59.
We will go through the whole manuscript to always make clear what the subject of discussion in the respective part is.

No information about the variations in the initial conditions of both the GCMs and RCMs is provided.
We did not include this in the original manuscript as it is already documented in the cited literature. According to the reviewer's suggestion we now summarize the initialization techniques in the revised manuscript.

Some specific comments:

14: Suggest: "Simulated inter-annual variability is evaluated against the observational dataset E-OBS and potential future changes under increasing atmospheric greenhouse gas concentrations are compared across the ensembles."
we will change this sentence

15: Delete sentence "To the knowledge of ..."
Changed accordingly

34: "Uncertainty of future climate projections can stem from at least three sources ..."
Changed accordingly

37: In L34, you mention the uncertainty in projected changes due to internal variability. Here you refer to "projected changes in uncertainty" understood as "projected changes in inter-annual variability" which addresses a different aspect of internal variability. These latter projected changes are subject to uncertainty due to internal variability as any other considered variable.
As mentioned in the general comments at the beginning, we will clarify our use of terminology and how they are connected more properly; especially the differences between internal variability, stemming from initial conditions and the inter-annual variability; this includes the usage of "uncertainty" as in this case

55: Using IMV to quantify IAV should not be motivated by "convenience" but by an advantage. What is the advantage here? Disturbing low-frequency variations are said to be small for seasonal mean and heavy precipitation. What about temperature? Using e.g. a running standard deviation over detrended 30-yr periods would not be sensitive to low-frequency variations. Further, it would be calculated over the same period (30 years) instead of over 16-50 years. IMV is similarly prone to biases due to events in the external forcing.
The advantage of using IMV as an estimate for unforced IAV is particularly relevant in the presence of non-linear forcing. For instance as pointed out by reviewer #2 the response to a volcanic eruption cannot be easily separated from unforced IAV. Also the anthropogenic forcing since 1950 has not been linear in time. IMV is an elegant way to get around this challenge. Since we use standard deviation as a metric it is also not sensitive to the use of 16, 30 or 50 years. The approach of using IMV as an approximation of IAV has been used in several previous studies, e.g. in Leduc et al. (2019, p. 681), where the authors state that "In the case of a climate system under transient forcing, the use of Eq. (1) to assess temporal variability using the inter-member spread involves weaker assumptions than calculating the residual temporal variability from detrended time series.", based on a study by Nikiéma, Laprise et al. (2018) [DOI 10.1007/s00382-017-3918-0].

63: "However" doesnlt make sense here.
removed it

72: I guess you mean they found significant changes in inter-annual variability only in a small number of CMIP5 models.
Exactly. We will change the sentence to make it clear.

118: I assume "surface temperature" refers to 2-m air temperature and "precipitation sums" to accumulated precipitation. Pleas clarify.
In two ensembles the variable is specified as "near-surface air temperature" and in one ensemble as "2m air temperature" – we will change the terminology to be more accurate. Although the term "precipitation sum" seems to be as unambiguous as "accumulated precipitation", we will change that.

121: Analysis is not limited to summer. A heat wave in winter, though, does not have obvious societal impacts.

The heatwaves are here defined as consecutive days above the 95th percentile calculated across all days of the year and not as time varying percentiles as in some other definitions. We can expect that all these days occur during the summer months. Experiencing this during winter seems very unlikely, even under RCP8.5. And even if so, such high temperatures (assume a summer heatwave as seen in the current climate) would probably have even worse impacts when occurring in winter, especially for ecosystems.

140ff: A reference to Fig.5 is missing.
added

145: "normally" distributed might be more appropriate than "randomly" distributed. The latter more suggests an equal distribution.
true, changed accordingly

159: Why detrended by the ensemble mean rather than by each member individually? The trends are subject to internal variability at lower frequencies and can influence the calculated inter-annual variability.
We aim at estimating the total unforced variability at interannual time scales including the contribution from low frequency variability. When removing a linear trend from each individual realization we remove some of this variability e.g. if the first or last year is extremely warm or cold. The multi member trend is a much better estimate of the forced response than each member's trend.

164: IMV is only insensitive to trends if the trends are the same among the ensemble members. And it is not insensitive to external forcing effects. E.g. if the variability of a specific variable is significantly lower after a volcanic eruption, the IMV would decrease as well. In fact, I would expect the IAV to be generally larger than the IMV (Fig.2). Any idea why IAV < IMV?
The first idea we had was that it is because the IAV data is detrended, while the data for IMV is the original values. However, after looking at the methodology again, we found that both IAV and IMV are based on the ensemble-mean-detrended time series for this comparison. Right now, we cannot explain why there is a systematic bias towards higher values of IMV, but we hope to find some explanation for the revised manuscript. This will be part of the extended discussion on the similarities and differences between IAV and IMV.

218: The E-OBS time series might also be too short to infer a representative pdf, in particular for extremes.
We will generally emphasize the limitations of an observed reference in this context, as it is always just one realization

229: Accronyms such as IMV are not used consistently.
Changed accordingly

234ff: Are these changes significant? For green and blue, the end of the tas-DJF time series shown in Fig.8, for example, seem to be close to or even within the historical ranges shown in Fig.2. This means that the future ranges clearly overlap with the historical ranges. Resting a significance test of a linear trend on smoothed time series, as done in Fig.9, is not valid.
Thank you very much for this useful comment. We took it up and totally changed the analysis of future changes in IAV (former figures 8 and 9). The new methodology, based on Brown-Forsysth tests (which is less susceptible to non-normality than a regular F-Test) for equal variances is introduced now:
Overlapping 30-year periods that are shifted by one year each are the basis (1961-1990, 1962-1991, …, 2070-2099) to detect changes in the variance of annual data (detrended by the ensemble mean) for each indicator. A Brown-Forsythe Test on equal variances (alpha=0.05) is performed for each of the periods with respect to the first period that serves as a reference (1961-1990). This is done for

*each member individually. We thus get information on changes in the variance for each period and each member. Together with the sign of change in variance in each of these cases, we can extract the number of members per period that show a statistically significant change (positive or negative) in variance (*Figure 1*). Results show that only a minority of members shows significant changes of variance. For the example of tas-DJF (ME), all members show a decrease in IAV, calculated as standard deviation of detrended 30-year periods (*Figure 2*). However this relatively clear change in the metric "standard deviation" does not imply significance in all members, as you stated correctly. This is what we can now show with the improved methodology, and this also exceeds much of the analysis in existing literature. Here, often no significance test of detected changes in inter-annual variability is performed. Many studies just inform about the robustness of change (e.g. by stippling in maps), measured by the accordance of (usually) 67% of the models of multi-model ensembles. This does however not allow information about the significance compared to a reference climate.*

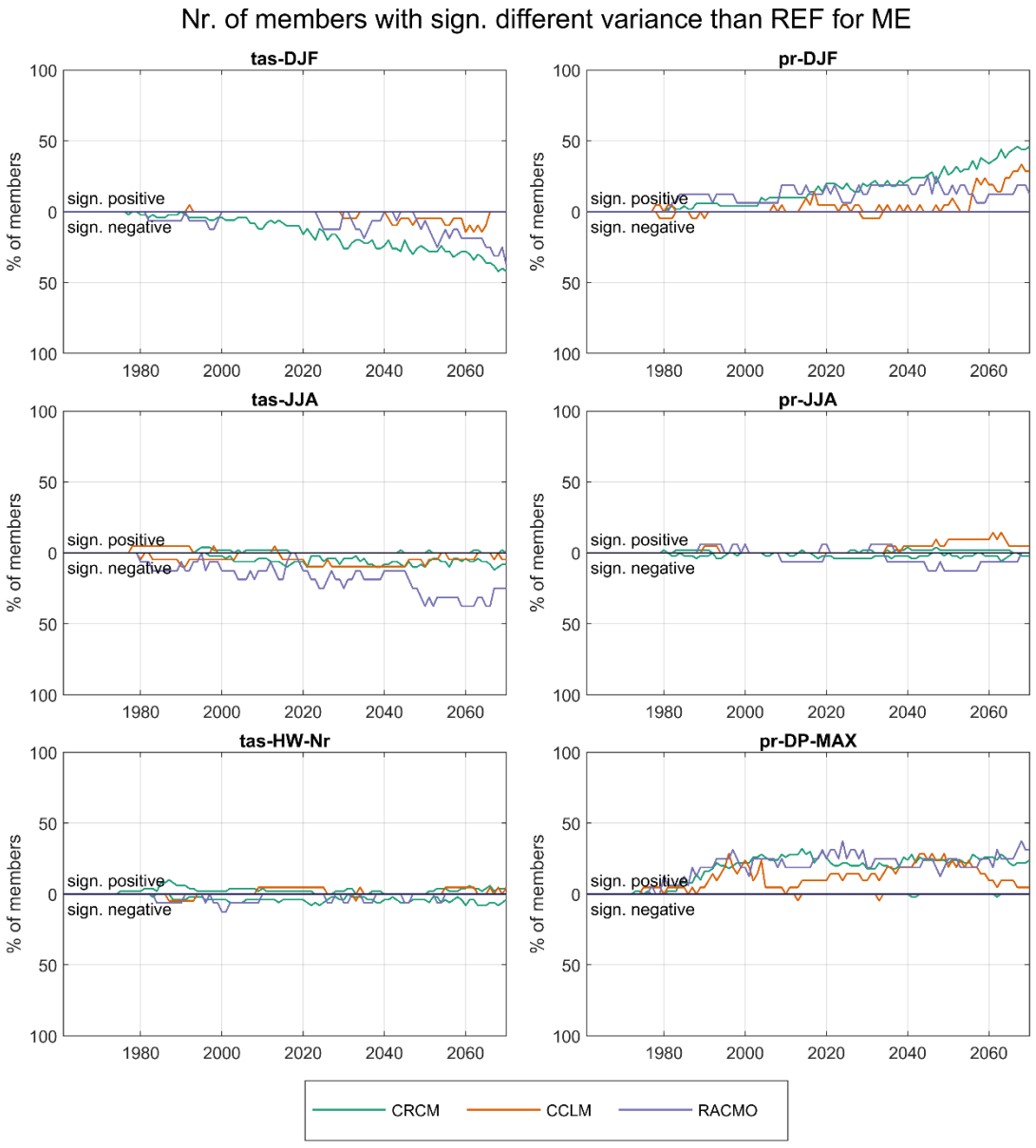

Figure 1: Percentage of members showing a significant change of variance in the respective period against the reference period 1961-1990 for the region ME. The x-axis depicts the starting year of each 30-year period. Positive changes are shown upwards, while negative changes are shown downwards.

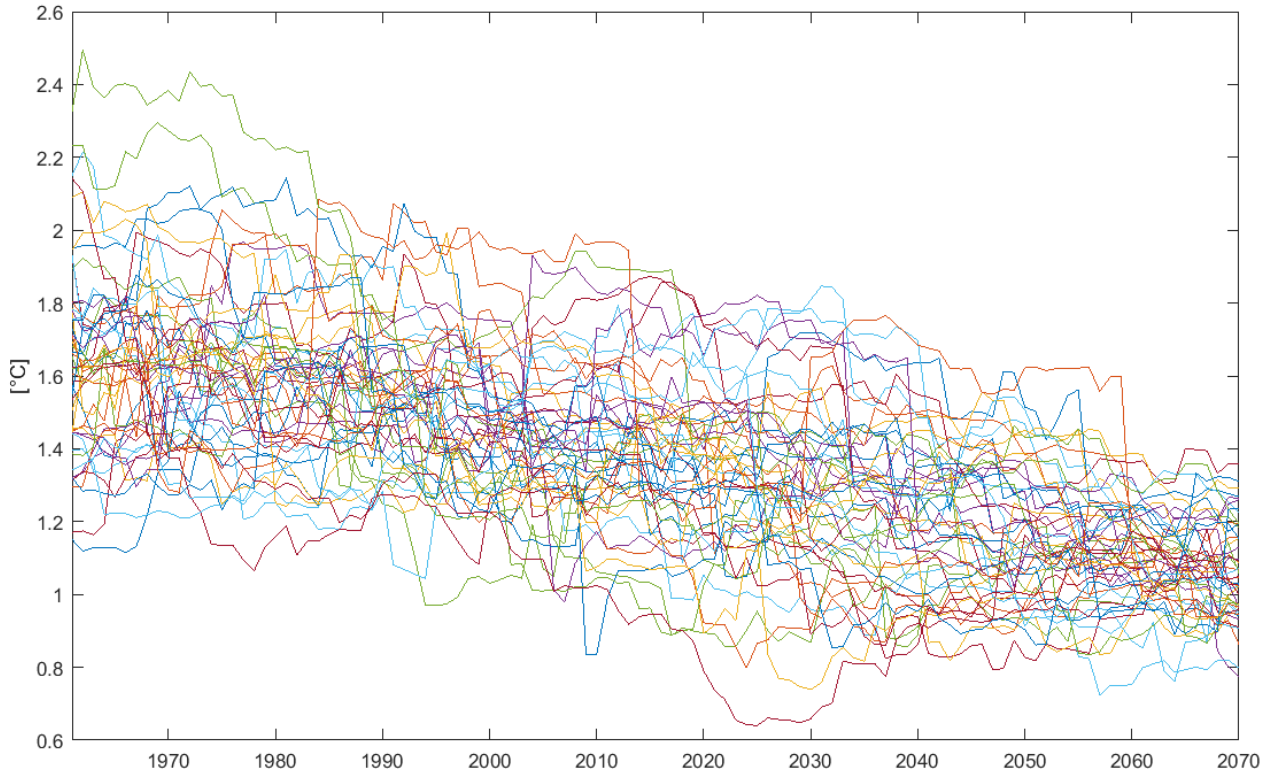

Figure 2: Temporal evolution of IAV (measured as standard deviation across 30-year periods, detrended with the ensemble mean) for the 50 members of CRCM fir tas-DJF in ME

Additional to this analysis of changes in IAV itself, we still think the concept of IMV as an approximation for IAV is useful, as it can incorporate both the information on the general direction of change as well as the significance of these changes. We therefore perform a similar methodology as for IAV on the IMV data: in contrast to IAV, IMV gives one value per year that we can easily plot for each SMILE. We then apply the Brown-Forsysth Test on the variances of all members per year to detect significant changes in the variance. This is done for all years with respect to the variance of the first year (1961). The results are shown in Figure 3 for ME. In only a very few cases, the IMV change is significant (solid line type). Why the IMV shows increasing tendencies for tas-JJA, while there are a rather negative tendencies in the IAV (Figure 1), needs to be clarified for the revised version. It might well be that this inconsistency exists because the IMV is based on the original data, not on detrended time series as used for the calculation of IAV in Figure 1. Unfortunately we were not able to perform the analysis with detrended data for the IMV data of Figure 3 until the deadline for this response. We will handle this in the new chapter, where all issues concerning the connections between IAV and IMV will be discussed.

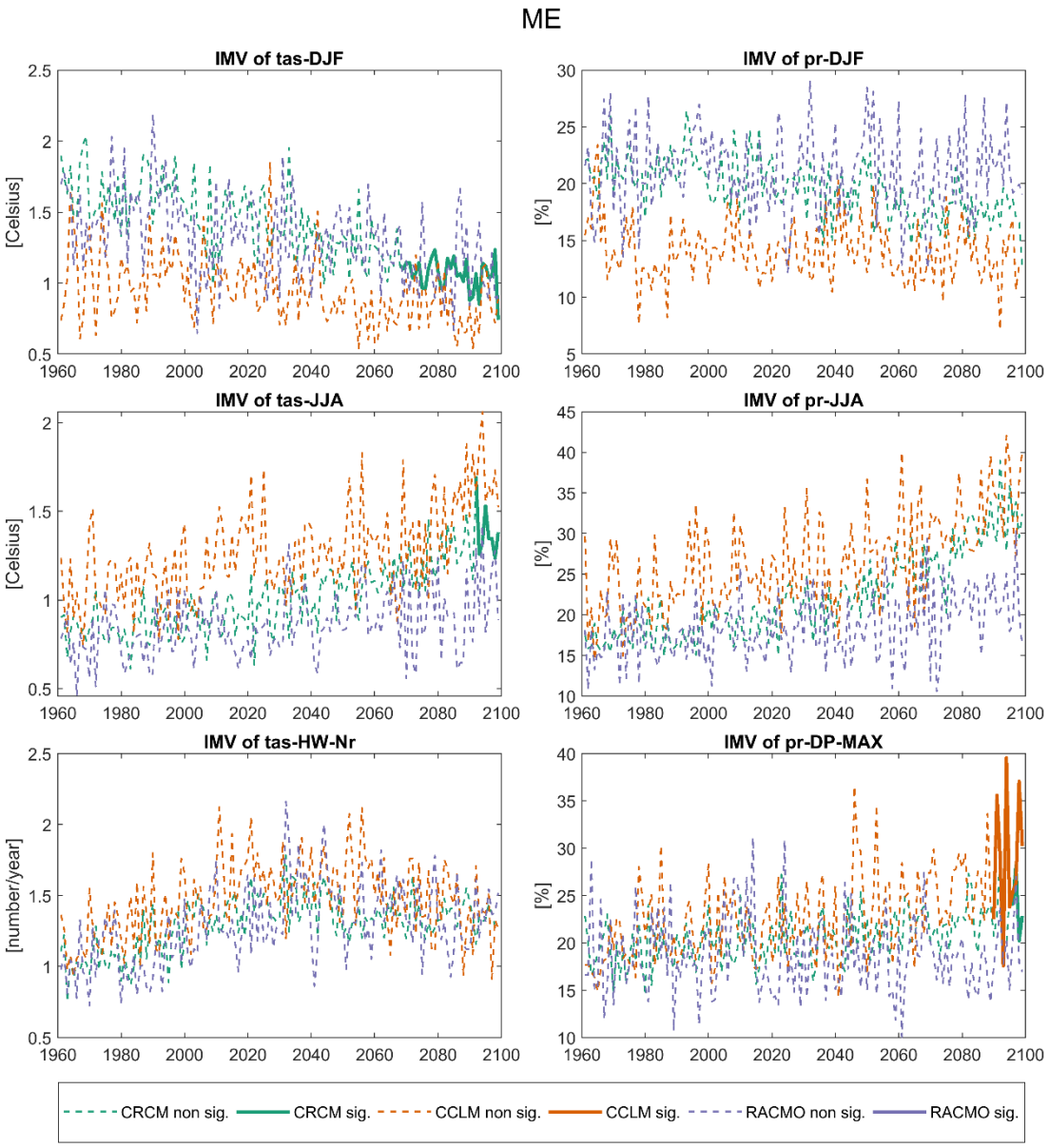

Figure 3: Temporal evolution of IMV per year. For each year the significance of change is evaluated with a Brown-Forsyth Test on equal variances against the first year (1961). The line style changes from dashed to solid when all following years show a significantly different variance, indicating a consistent change in variability.

254: Ensemble means are not shown in Fig.5.
correct, we will add lines to show the ensemble means

258: The correlations shown in Fig.S10/S11 only reflect the signs of the respective changes shown in Fig.5/S2/S4 and do not add any information. In fact, a correlation analysis between time series subject to trends is heavily influenced by the trends and thus not quite meaningful.
These figures and their discussion will be removed

272: Scientific discussions are always critical.
of course

285: Many biases might be inherited from the driving GCMs. A comparison is highly recommended.
As mentioned above, while a comparison is highly interesting, it is outside the scope of the current manuscript.

290: What is the "coefficient of variation" applied by Giorgi et al.? Why not using it?
The coefficient of variation is defined as the ratio of std/mean, which is the same as using the standard deviation of relative differences from the mean, as we did; so the result is the same, although slightly different calculations were performed. We mention the work by Giorgi et al. to add a reference to similar thoughts around the interpretation of variability for precipitation based indicators. This will be clearer in the revised manuscript.

294: "Agreement and dissent" evaluates the results as kind of ambivalent. This does not fit with "even better agreement" at the beginning of the next sentence.
we will improve the wording in this section by discarding the "even"

304: If I understand the approach correctly, from a future increase in IMV one cannot infer whether this increase is due to an increase in inter-annual variability or due to an increase in the spread of the mean states caused by internal variability. In L55 it is said, that it is valid to use IMV as an approximation for IAV if long-term variations are small compared to IAV. However, long-term variations (including the inter-member spread in the projected change signals) need to be compared with the projected changes in IAV, not only with absolute IAV.
Note that ideally IAV in transient simulations could be calculated over a very long time period of e.g. more than 100 years and thereby it would also include a contribution from low frequency variability. Note that we do not explicitly use a high-pass filter here to calculate IAV. However, we here calculate IAV over a detrended 30-yr period. Thereby our calculation of IAV does not account for low frequency internal variability, which is why we stated that it is valid to use our IMV calculation as an approximation of IAV.

Likewise, the change in future IMV may arise from changes in low-frequency variability (which would lead to a larger spread in 30-yr means) or high-frequency variability e.g., IAV within 30 year periods. Thus we compare in the revised version changes in IMV and IAV calculated over 30-yr periods.

319: Why is it plausible that the statistics of the length of dry periods increase for RCP8.5? In northern Europe, precipitation is projected to increase due to the enhanced moisture transport from low to high latitudes.
Increases in the length of dry periods do not necessarily contradict increasing precipitation in general; CRCM5 also shows an increase for the ensemble mean of pr-DP-MAX for Scandinavia from 18 to 20 days, although the IMV is quite large (6-8 days), why this change is probably not significant. However, EURO-CORDEX data from Jacobs et al. (2014) do not show significant increases for the length of dry spells in Northern Europe. We will change this sentence accordingly, also differing between heatwaves and dry spells.

329: I highly recommend to include the RCM-GCM comparison in the present study. Whether downscaling with respect to inter-annual variability is important or not can only be demonstrated by such a comparison.
see above

338: I disagree. An increase in inter-annual variability does not imply an increase in the uncertainty of climate projections due to internal variability. Climate change signals are typically based on climatological means. The spread of these is referred to as uncertainty due to internal variability and this metric does not necessarily depend on inter-annual variability.
As stated above, we will better disentangle the two terms and adapt the conclusions drawn. We agree that an increase in IAV does not necessarily imply an increase in uncertainty of climate projections due to internal variability.

340: The mean is not shown but required to assess this statement.
added