<u>Response to Reviewer #2</u>

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.

The authors use large ensembles to compare the representation of internal variability in three regional climate models forced by historical and a future scenario forcing. They use observation-based data as a benchmark for the historical period. Large ensemble simulations of single climate models are an essential tool for estimating uncertainty of climate change projections due to internal unforced variability, for detection and attribution studies and so forth. The present study is therefore useful as a validation of such tools and to enhance our understanding of unforced internal climate variability at regional scales. My main concerns with this work however are its presentation, which is confusing at times, the implementation and interpretation of the methodology, and the interpretation of results.

<u>Major:</u>

1. The authors recognize that there is confusion in the literature on what is meant by “internal variability" of the climate system (Lines 44-46). I agree. I also agree with the authors’ definition of internal variability (Lines 45-51, although it can be shortened). However, in many occasions the authors seem to equate internal unforced variability with inter-annual variability, which add to the confusion (e.g., Lines 10-12, Lines 53-55). I would recommend to clearly define the two from the onset noting that internal variability is unforced whereas inter-annual variability can be externally forced by natural and anthropogenic aerosols, GHGs, solar radiation, land change use and so forth. If inter-annual variability is understood as derived from detrended time series, then explicitly say so from the onset, and clearly indicate how they are detrended.
We agree that IMV of annual data might differ from IAV in the presence of forced changes due to aerosols, GHGs, solar and volcanic forcing. In the revised manuscript we will state more clearly that IMV of annual data is used as (a very good) estimate *of IAV. Indeed, we determine IAV in observations and simulations as the standard deviation of detrended time series. We will clarify the terms and their usage in our study throughout the whole manuscript. We took the approach of using IMV of annual data as an approximation of IAV from 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]. In the current version of the manuscript we show that IMV is indeed a good estimate for IAV in the reference period. In the revised manuscript we will also include a comparison of future changes in IMV of annual data and IAV calculated from 30-year periods.

2. The methodology presented in lines 140–150 is used to asses the interannual variability in the models against that in observations (results in Fig. 5-7). It should be clearly stated that this methodology is not an assessment of model internal (unforced) variability alone, since the time series are affected by the forced signal. Therefore, if there is no agreement between model and observations, we should not conclude that the model representation of internal variability is incorrect, as it may be consequence of the externally forced signal (e.g., the model may have a perfect representation of internal variability, but a too strong response to volcanic eruptions leading to disagreement in the anomaly distributions of Figs. 5-7). On the other hand, I would agree that if the observed and modelled distributions are coincident, this would suggest that both the model response to external forcing and its internal unforced variability are well represented. I don't think this point is clearly made in the methods section and the discussions of sections 4.2 and 5. The way the methodology is presented and the results discussed seem as if the model response to external forcing and that from the observations are in perfect agreement.
We will take up these valuable comments in the methods and discussion section. As written above, we plan to include a section on the biases of the models. These comments can be an integral part of this extended section.

3. Based on what is expected from the methodology introduced in lines 140–152, the distributions for the ensemble members in, say, Fig. 7 should largely coincide. They don't. In some cases they are quite different as noted by the authors. It is unclear then how to assess the agreement between model and observations based on these distributions. Are these differences because of a small sample size, or because the ensembles are not large enough? Could they be consequence of poorly sampled (multi)-decadal variability? Can the authors comment on this? I didn't quite follow the rationale of the last sentence of section 4.2, particularly the bit about added "information".

The fact that different members disagree on the distribution of the annual anomalies (internal variability) is exactly the reason why large ensembles are required to accurately estimate inter-annual variability and changes therein. We expect the differences indeed to mainly originate from the sample size (30 years), but *low frequency variations that differ for the 30 years of the reference period between members, with which the members are normalized, may play a role as well. The last sentence of section 4.2 tries to comment on the occurrence of these outlier members. They have a distinctly different distribution than all other members, just because of their initial conditions (this is meant by "how large the influence of internal variability can be…"). The statement on "added information" points to the question of the size of SMILEs needed to get a satisfying quantification of internal variability (thus, how much results differ from member to member). These outlier members can "add information" on the magnitude of internal variability even after 49 relatively similar members. Let's pretend you take a look at these distributions one member after another and the outlier is the last one. You would guess that the distributions of the 49 members you had seen so far give you a good approximation of how further members' distributions would look like. However, the last one looks quite different, thus "adding information". We will make this clearer in the revised manuscript.*

<u>Minor:</u>

In the title: Consider changing “variabilities" to “variability"
changed accordingly

Line 11: “... (here: inter-annual variability) ...". Do you mean “on inter-annual timescales"? Inter-annual variability is affected by both externally forced and internally unforced variability. See comments above.
We will rephrase this sentence of the abstract according to a generally better definition of terms.

Lines 53-57: I don’t think this is accurate and should be reworded. The ensemble spread about the mean can be used to measure the internal unforced variability of the model, but may not be representative of inter-annual variations in the presence of, say, a strong volcanic eruption. Therefore, using IMV to assess IAV may be a good approximation in some cases, but may also be in error. This should be clearly stated.
We will more clearly state that IMV on annual data is an estimate of IAV in the revised manuscript.

Line 105: I believe the work by Fyfe et al., 2017 uses the regional climate model CanESM2-CanRCM4 which is different from CanESM2-CRCM5.
Fyfe et al was used as a source for information on the driving CanESM2 data, describing the initialization and creation of members. More recent papers seem to be using Kirchmeier-Young et al (2017) [10.1175/JCLI-D-16-0412.1] for a description of the CanESM2 LE, so we will exchange the source.

Line 115: Although the authors provide a reference, I would find useful a brief comment on the weakness of the E-OBS dataset.
We will include a short comment.

Line 120-124: Consider moving the text “These indicators (...) transport of rivers and many more", to the introduction and leave this section only for the methods.
Our general plan for the revised version is to adjust the storyline of the paper more to an RCM users’ and impact modelers’ perspective. We will therefore extent the part on impact-relevant information and, move them to the introduction and pick them up in the conclusions part.

Lines 129-139: The discussion on whether the indicators should be computed on the original grid to evaluate the averaged quantities, instead of regridding first and then evaluate the averages over a common grid, seems too long. The authors claim that both approaches give similar results and chose the former, which I believe is the recommended approach (Diaconescu et al., J. Hydrometeorol. 16, 2301–2310). The discussion could be shortened, and this reference cited.
Thank you for the hint to this publication.

Line 140–150: It would be helpful to have explicit references to Figs. 5-7 when discussing the method. Expressions like “we then plot" or “are plotted" are used without showing an actual plot.
we will add references

Line 143: Change “neglected" by “avoided" or the like.
we will

Line 165: The authors claim that the spread between members is a well suited metric to examine projected changes of inter-annual variability. I would not agree, unless it refers to detrended inter-annual variability (i.e., after removing the ensemble mean). If so, please clearly specify, although note that the previous sentence (line163) would then be confusing because both IMV and IAV should be insensitive to external forcing, as no temporal autocorrelation is assumed (line 58).
The time series are indeed detrended with the ensemble mean as stated right before in line 159. In

the revised manuscript we will more clearly define IMV and IAV and their characteristics, as discussed above.

Figure 3 and 4: I'm not sure how to interpret the changes in precipitation. It seems that the plots are intended to show changes in the climatological mean for each ensemble member over two different time periods. Why not showing just that? Precipitation is given here in %. Are these differences of relative values, or rel- ative values of differences? Line 171 defines relative indices by dividing the en- semble anomalies about the mean by the ensemble mean. I can see how this may be useful to assesschanges in ensemble spread, but this is not what is shown in Figs. 3 and 4, right? Please define clearly the quantities in these figures and clearly describe their behaviour.
You are right, they just show the differences between the mean states of two periods. The change is given in % for each member, thus an increase from 100mm to 120mm will result in a value of 20% for that member. The definition from line 171 is referring to the change in internal variability in the future (Fig. 8). We will add more information to the figure description, and in the methods part, where it is indeed missing at the moment.

Line 202: Remove "years 1957–2015 in" from the parentheses.
probably redundant, yes

Line 207: "(slightly) too low" is odd.
we will change the wording here

Line 209: What does the authors mean by density functions "somewhat inflated"? Please clarify.
The comparison between observations and simulations is based on the original data (opposed to detrended data). The total variability shown here thus includes the inter-annual variability around the mean (IAV is defined as the variability of detrended time series) and the deviation due to a trend in the mean climate state, i.e. the variability could be somewhat larger than IAV calculated from detrended time series.

Line 305: Change to "an statistical model" instead?
that is shorter indeed