Response to reviewers' comments on

esd-2021-8

"Balanced estimate and uncertainty assessment of European climate change using the large EURO-CORDEX regional climate model ensemble"

20 September, 2021

We thank the Editor, the two reviewers for their constructive comments, as well as the additional comments by Rasmus Benestad and Richard Rosen. These comments are reported in *blue* and in italic font. Most of the suggestions have been taken into account and we hope that the revised version will answer the concerns raised by the reviewers. Please see our response to the different comments below.

CC1: Rasmus Benestad

CC1#1: *Evin et al. present an impressive and extensive study involving the ANOVA method and data augmentation to estimate balanced mean change with an approach called 'QUALYPSO' and to distinguish between sources of uncertainties: RCPs, GCMs, RCMs and internal variability. This is a contribution to progress in terms of understanding downscaled model projections and the use of ensembles.*

Thank you for this positive comment.

CC1#2: I nevertheless have some suggestions concerning the background and introduction to this study. My impression is that there are many papers presenting RCM results that ignore related work done through empirical-statistical downscaling (ESD). I think that acknowledging such work in many cases would strengthen them. Here for instance, the statement "the largest MME of regional climate projections ever produced" is not quite correct since ESD efforts for a while have produced larger Multiscenarios Multimodel Ensembles (MME). For instance, a total of 254 downscaled simulations, each for 200 years, provided a basis for new ways to present large ensembles in Benestad et al. (2017; DOI: 10.1016/j.cliser.2017.06.013).

Thank you for this comment. We totally agree that ESD approaches have their merits and are a valid alternative to dynamic downscaling approaches provided by RCMs. This was indeed missing in the first version of the manuscript and has been added in the discussion (see paragraph "MME projections of future regional climate") and in conclusion.

Concerning the introduction and the reference to *"the largest MME of regional climate projections ever produced"*, we now precise that we are here referring to “the largest MME projections based on regional climate models ever produced”.

CC1#3: I would also argue that any effort to provide a robust estimate of total uncertainty in connection with downscaling should involve both RCMs as well as ESD, because these two approaches have different strengths and weaknesses independent of each other. They draw on different sources of information. For instance RCMs may be biased because of inconsistencies with the driving global climate model (GCM or ESM - here I use GCM referring to both). E.g. using different parameterisation schemes than the driving GCM, or there may be differences in outgoing longwave radiation (OLR) at the top of the atmosphere because the RCMs produce different rain/cloud climate to the driving GCM. Furthermore, both RCMs and ESD rely on the link between large and small scales being stationary, but in different ways: ESD in terms of the calibration of historical predictors and predictands; RCMs in terms of their parameterization schemes that provide a large-scale aggregation of unresolved small-scale processes.

Thank you for this comment. We are aware of the strengths and weaknesses of ESD and RCM approaches. In addition, we agree with you that complementary between dynamical and statistical approaches is a key to provide useful regional climate information. In our opinion, the systematic inclusion of projections based on ESD can be discussed. Indeed, bias correction techniques or data augmentation methods can also be relevant depending on the application. In the case of our study, we now underline the complementarity between statistical and dynamical approaches and a new sentence has been added in conclusion: “*In addition to the dynamical downscaling approach provided by regional climate models, several alternative methods are available in the literature in order to improve the uncertainty assessment of future climate change (e.g., empirical-statistical downscaling methods, Gutiérrez et al., Benestad et al., 2017a)*”. A detailed discussion of the strengths and weaknesses of ESD and RCM approaches seems to be out of the scope of the study.

CC1#4: A comment on "the Arctic warming amplification and to the regional snow-albedo positive feedback" is that the Arctic amplification is even more pronounced at the high latitudes, during winter, when there is no sunlight (during the polar nights). It's not so obvious that this is due to an albedo effect because it’s dark (and there is also often a cloud-cover present).

Concerning the Arctic warming amplification, the full sentence in the original manuscript is “*In winter and with the scenario RCP85, NEU and CEU are warming substantially more than the MED area likely due to the Arctic warming amplification and to the regional snow-albedo positive feedbac*k”. This is a misunderstanding here, both processes (Arctic amplification and snow-albedo feedback) are mentioned as 2 possible processes able to explain the extra warming in North and Central Europe (NEU, CEU) but they are not linked in our phrasing. In particular, we confirm that the snow-albedo feedback can play a role over land in Central Europe (CEU) in Winter.

CC1#5: It's a bit surprising that the internal variability converges to zero at 2100 as it seems like in Figure 2 - column 5 when the total temperature change only is a few degrees C (especially for northern Europe). I suggest checking the calculations. Also see Deser et al. (2012; DOI: 10.1038/nclimate1562).

Internal variability of a modelling chain at a given time typically refers to the variance of all realizations that could be obtained with that chain at that time from multiple runs. In our work, it is estimated as the (temporal) variances of 30-years average deviations from the estimated climate change response of the chain, and the internal variability estimate of the multimodel EURO-CORDEX ensemble is estimated as the multichain mean. It is thus assumed to be constant over time and does consequently not converge to zero.

For mean temperature changes in winter, for region CEU, internal variability is equal to 0.023 [°C^2] as a variance and 0.15 [°C] as a standard deviation. At the end of the century, the total uncertainty variance (including internal variability) is equal to 1.86 [°C^2] as a variance and 1.36 [°C] as a standard deviation. This large total uncertainty variance is explained by the significant differences obtained between climate change responses for the different RCPs (in dark green in Fig. 2, column 5) and for the different GCMs (in dark blue in Fig. 2, column 5). At the end of the century, the contribution of internal variability (as a **variance**) to the total uncertainty (again as a variance) is thus very small (around 1.2% of the total variance). This small contribution corresponds to what was presented in column 5 of Figures 2 and 4 of the original manuscript.

As mentioned in the manuscript, the estimate of internal variability would have been larger for 20-yrs, 10-yrs or 1-yr averages (see l. 218-225 of the revised manuscript) but it would not have been larger than 2 times for 20-yrs averages and 3 times for 10-yrs averages because of the expected temporal correlation in 10-yrs averages (likely induced by internal variability), see the additional results in response to the comment RC1#6 below.

It is difficult to compare our results to those presented in Deser et al. (2012), the data and the statistical framework used for the estimation differing significantly: estimates from Deser et al. 2012 are obtained from a 40-member ensemble of 57 years (2003-2060) obtained with one GCM for one single emission scenario (SRES A1B) and internal variability is estimated from the standard deviation of 8-year low-pass filtered data where the low-pass filter is a 5-point binomial filter. Internal variability can vary significantly from one model to the other and as mentioned above, internal variability also depends on the aggregation scale of the studied variable (8-yrs aggregation in Deser et al. 2012 compared to our 30-yrs aggregation). Even with these differences, the internal variability deviation estimate in Deser et al. 2012 is coherent with ours: it varies in Europe from below 0.3 (Southern, Central) to below 1.2°C (Scandinavia) in DJF and is below 0.9°C (below 0.3 in Scandinavia) in JJA (see Fig. 16 In Deser et al. 2012).

We have added some additional comments in Section 4:

"*For both variables, internal variability being considered constant over time, its contribution to the total variability dominates during the first decades and rapidly*

decreases due to the larger variability at the end of the century. This moderate contribution can also be explained by the fact that internal variability is obtained from 30-year averages of seasonal temperature and precipitation values here. A larger contribution of internal variability for smaller temporal aggregation scales is illustrated in Figures S5 and S6 of the SM, using 1-year, 10-year, and 30-year aggregation scales, for the 3 SREX regions in winter. At an annual time scale, the contribution of internal variability for relative changes of precipitation in winter is up to 80% of the total variance in the Mediterranean region in 2100 (Fig. S6). For temperature, the contribution is smaller but reaches 40% in CEU at an annual time scale in 2100 (Fig. S5)."

and in Section 6

"*At the end of the century, internal variability is small, especially for temperature changes. It only exceeds 10% for precipitation changes over specific regions (e.g. South-West of Europe in winter, North Africa in summer). As mentioned previously, the internal variability variance and then its relative contribution to the total variance depend on the temporal and spatial aggregation scales considered for the studied variable. As shown in Figs S1 and S2 in the SM, the internal variability relative contribution is much stronger for shorter temporal scales*".

CC1#6: GCM uncertainty over sea may be a result of incorrect sea-ice cover since the temperature shoots up in the air where it retreats, as mentioned in the discussion. This has been interpreted as a known shortcoming in the past, and it's a bit surprising if the models with sea-ice in this region also are considered among the most trustable CMIP5 GCMs concerning the wintertime sea-ice cover. I suggest checking this.

GCM uncertainty over the sea can be related either to the SST change pattern or to the sea ice cover (SIC) pattern indeed. We agree that the presence or absence of sea ice can strongly modify the near-surface atmosphere temperature and therefore explain locally the wintertime GCM uncertainty. However, SIC can play a key role without invoking incorrect sea-ice cover. Indeed even with "good" GCMs, the uncertainty in the sea ice cover response is expected to be strong. To carefully check this point, we would need to access the SST and SIC of both the GCMs and RCMs. To our knowledge, this information is not available for RCMs. This said it is still possible that GCMs with implausible sea ice cover behavior are part of the EURO-CORDEX driving GCMs as we are not aware that the driving GCMs have been selected specifically on this criteria. This is however planned for the selection of the CMIP6 GCMs.

CC1#7: Uncertainties should probably not exclude the possibility of tipping points. In this case, it could be a reversal of the thermohaline circulation.

We fully agree with the reviewer. GCM simulations with tipping points should not be excluded a priori if they can be considered as plausible futures. It is not certain however that "reversal of the thermohaline circulation" can be considered plausible during the 21st century. Nevertheless, QUALYPSO method relies on the existing GCMxRCMxRCP matrix. It means that it can not be used to extrapolate towards GCMs that have not been downscaled by any RCM in the initial matrix. So to include GCMs with tipping points in

the QUALYPSO uncertainty range, we need to have them chosen by EURO-CORDEX. This is a limitation of the study and we have added the following paragraph in the discussion:

"*In addition, the QUALYPSO method, as most of the matrix filling methods, relies on the existing MME. In our case, it means that it can not be used to extrapolate towards GCMs that have not been downscaled by any RCM in the GCM x RCM x RCP matrix considered in this study. In addition, the results shown in this study are unavoidably impacted by the shortcomings of the climate models.*"

CC1#8: One suggestion: when it comes to precipitation, two key parameters are also the wet-day mean precipitation and the wet-day frequency. They are useful because they provide more actionable information than just the seasonal totals (their product with the number of days is the total precipitation amount).

Thank you for this comment. We agree that seasonal totals of precipitation are a very restrictive summary of the information provided by the climate projections available at fine temporal scales (e.g. daily). The same comment applies to temperature, seasonal averages being of limited interest for the most critical effects of climate change (e.g. intensification of heatwaves). 10-20 climate indices are customarily studied (see, e.g., Dosio et al., 2016). These analyses are obviously of interest and we expect to carry out such additional studies with QUALYPSO for part of them. Note that, however, for indices related to high quantiles, or involving a threshold, a bias-correction step is likely unavoidable as these indices represent characteristics of the statistical distribution that are typically poorly represented in raw climate projections.

CC1#9: Another suggestion is that methods from ESD can be used to study the connections between features provided by the driving GCM and the response simulated by nested RCMs. For instance, ESD calibrated with one GCM-RCM pair may be applied to a different GCM to compare with its RCM. This is a bit like 'hybrid downscaling'.

Thank you for this suggestion. We are well aware of the possibility to apply ESD-like techniques to develop hybrid downscaling. We however consider that this is out of the scope of the current study. The interest of SDMs and hybrid techniques to complete GCMxRCMxRCP matrices is now acknowledged in the revised manuscript, in conclusion ("*In addition to the dynamical downscaling approach provided by regional climate models, several alternative methods are available in the literature in order to improve the uncertainty assessment of future climate change (e.g., empirical-statistical downscaling methods, Gutiérrez et al., Benestad et al., 2017a)*".

CC1#10: Finally, ESD can be regarded as a way to test the uncertainties connected with GCMs and decadal variability, and results by Mezghani et al (2019; DOI: 10.1175/JAMC-D-18-0179.1) may seem to suggest that internal variability plays a bigger role on a regional scale than the GCM (is didn't use 30-year smoothing, however). These results also highlight the limitation posed by 'the law of small numbers'. One nice aspect of ESD is that it can incorporate a quality evaluation of GCMs.

Thank you for this comment that again highlights the potential interest of ESD and will be acknowledged in the future manuscript version (see the previous comment). Concerning the quality evaluation of GCMs, it is obviously another critical and key issue in climate impact studies. This issue is now recalled in the revised manuscript in Section 8 (subsection 8.3 "*MME projections of future regional climate*"). Indeed, the problem of the "law of small numbers" (the fact that few RCMs and GCMs runs are available in some MMEs) is only partially resolved by QUALYPSO by balancing the estimates. However, obviously, it cannot take into account information about GCMs or RCMs that are missing from the MMEs, and the constitution of the MMEs is thus a critical step, independently of the uncertainty assessment approach (e.g. QUALYPSO), see our response to the comment CC1#7 above. Concerning the internal variability, we recall here, as it was already stated in the manuscript, that internal variability strongly depends on the spatial (pixel scale, averages over SREX boxes, countries, etc.) and temporal (interannual variability, averages over 10-yr or 30-yr periods) scales, as well as the variable analysed (temperature, precipitation, etc.), see our response to the comment RC1#5 below.

RC1: Anonymous Referee #1

RC1#1: This is a comprehensive dynamical downscaling study of regional climate change in Europe. They applied a statistical approach called QUALYPSO to a very large ensemble of regional climate model simulations to partition uncertainties due to internal variability, model, scenario, etc..

We thank the reviewer for its interest in this work.

RC1#2: But the method to partition uncertainty has limitations (Hawkins and Sutton 2009), uncertainty here only means ensemble spread and it has nothing to do with errors. No observational constraints are taken into consideration.

We agree with the reviewer that our approach provides an estimation of the model ensemble spread, which is only one of the components of the "error" of the future projections.

Concerning the use of past observations to improve the final regional climate information provided, a new subsection "8.3 MME projections of future regional climate" has been included in the revised manuscript:

"*This study assesses changes between a future and a reference period, as most of the studies on this subject. The evolution of the climate from this reference period only is trusted and assessed, and not absolute values. However, past observations can also be used to improve the final regional climate information provided or to complement QUALYPSO-like approaches. For example, observations can be used to weight or to select GCM or RCM in the initial ensemble. Indeed, many papers question the "model democracy" approach and aim at estimating future mean changes and associated uncertainties by proposing different ways to combine the runs of the MME, mostly using weights (see Brunner et al., 2020, for a recent comparison). Those constraints are not*

applied in the current study but they could be considered in future works, as a complementary approach. Regional observations can also be used to correct a posteriori the regional climate simulations (Vrac and Friederichs, 2014) through statistical correction techniques. In the current study, QUALYPSO is not applied to postprocessed MMEs using such techniques but this could be easily done in future work by applying QUALYPSO after a correction step."

RC1#3: It's unclear whether the GCMs can represent internal decadal variability and its regional climate impacts reasonably well.

We agree with the reviewer that GCMs are known to produce very different internal variability from one GCM to the other (Deser et al. 2020), which tends to show that some GCMs overestimate/underestimate this characteristic of future possible climates. In this study, we acknowledge that QUALYPSO relies on the runs available in the MMEs with their limitations/drawbacks, including this possible misrepresentation of the internal decadal variability. This has been added to the discussion (Subsection 8.4):

"In addition, the QUALYPSO method, as most of the matrix filling methods, relies on the existing MME. In our case, it means that it can not be used to extrapolate towards GCMs that have not been downscaled by any RCM in the GCM x RCM x RCP matrix considered in this study. In addition, the results shown in this study are unavoidably impacted by the shortcomings of the climate models. For example, GCMs are known to produce very different internal variability from one GCM to the other (Deser et al., 2020), which tends to show that some GCMs overestimate/underestimate this characteristic of future possible climates. QUALYPSO relies on the runs available in the MMEs with their limitations/drawbacks, including this possible misrepresentation of the internal decadal variability."

RC1#4: Although the uncertainty partitioning method has been quite popular in literature, it will be helpful to add some discussions on this method.

There is already an extended paragraph at lines 58-71 of the current manuscript. We believe that this discussion is sufficient, and more details can be found in the previous study (Evin et al., 2019). Compared to existing uncertainty partitioning methods, QUALYPSO 1/ use the "time-series approach" to disentangle internal variability and the climate response, 2/ balance the estimates, 3/ has been applied to RCP / GCM / RCM MMEs on the contrary to many related studies.

RC1#5: Particularly, if the uncertainty here has nothing to do with error, how should we use the results to help provide robust estimates of future regional climate change.

First, it is important to remind here that the study assesses changes between a future and a reference period, as most of the studies on this subject. The evolution of the climate from this reference period only is trusted and assessed, and not absolute values. Second, as indicated above (comment RC1#2), QUALYPSO already applies observational constraints by assuming that the MME has been correctly built, and can be applied on bias-corrected projections (which was for example the case in Evin et al.,

2019). Third, many papers question the model democracy" approach and aim at estimating future mean changes and associated uncertainties by proposing a different way to combine the runs of the MME, mostly using weights (see Brunner et al., 2020 for a recent comparison). This study provides many insights about the different uncertainty components, models diversity in terms of climate change response. As such, this work should be considered complementary to the aforementioned methods. In QUALYPSO, contrary to most existing approaches, our estimates are "robust" to the subsampling of the complete MME (i.e. possible combinations of RCP / GCM / RCM). In future works, we could also consider using the weights provided by other methods to either select the most relevant runs of the MMEs or to propose weighted estimates of future mean changes and associated uncertainties.

Subsection 8.3 "MME projections of future regional climate" now discuss these aspects (see also comment RC1#2 above).

RC1#6: In addition, the study finds relatively small contribution from internal variability, this conclusion seems to contradict with many studies that emphasize the importance to run large initial-value ensembles. It would be useful to discuss whether this is caused by the analysis method here (e.g. whether this method tends to produce narrow internal variability uncertainty).

The small contribution of internal variability can be explained by the temporal and spatial windows. The relationship between internal variability and the spatial aggregation scale was mentioned at l.223-232 of the original manuscript, based on the results obtained for the different countries and their respective capital cities. Concerning the temporal aggregation scale, in Evin et al. (2019), the same method, when applied to annual values of total precipitation in the French Alps (see their Fig. 7), shows contributions of internal variability between 40% and 70%. In the revised manuscript (l. 218-225), this larger contribution of internal variability for smaller temporal aggregation scales is illustrated with additional figures in the SM, using 1-year, 10-year and 30-year aggregation scales, for the 3 SREX regions, in winter (see Figures 1 and 2 below). At an annual time scale, the contribution of internal variability for relative changes of precipitation in winter is up to 80% of the total variance in the Mediterranean region in 2100 (Figure 2). For temperature, the contribution is smaller but reaches 40% in CEU at an annual time scale in 2100 (Figure 1).

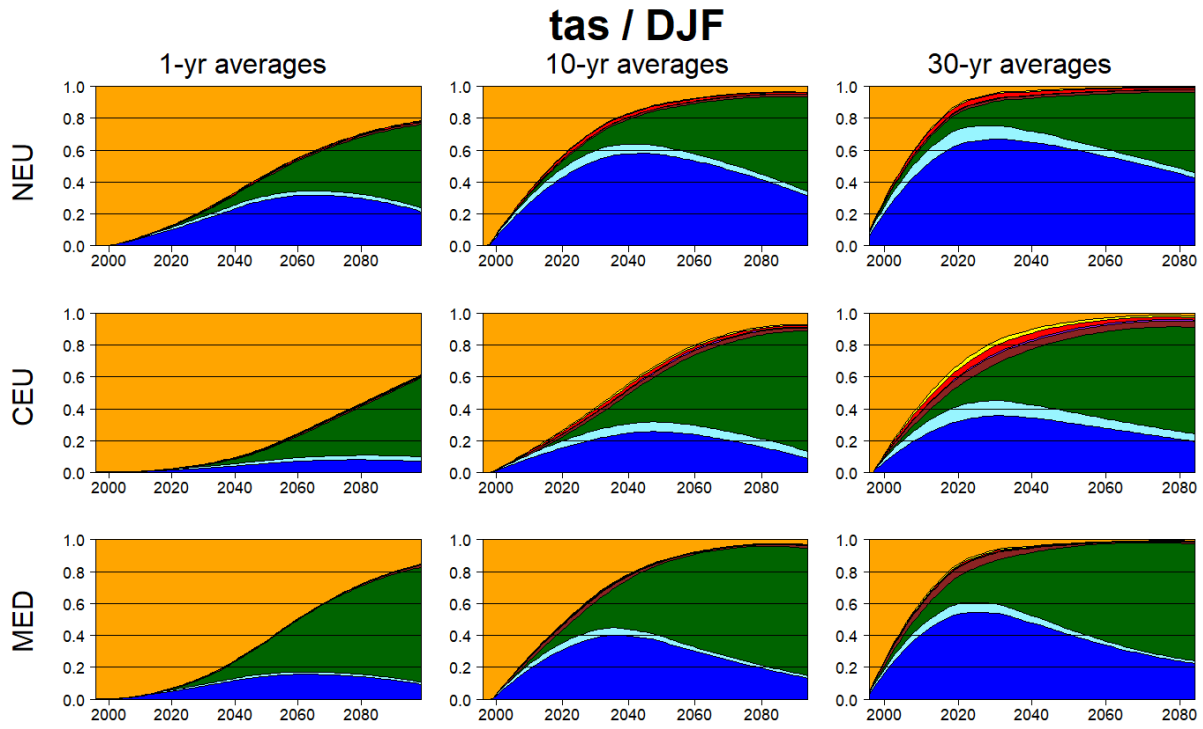

Figure 1: Fraction of total variance for absolute changes of mean temperature in winter (DJF), as a function of time, for different temporal aggregation scales: 1-year (left plots), 10-year (middle plots) and 30-year averages (right plots), for the 3 SREX regions.

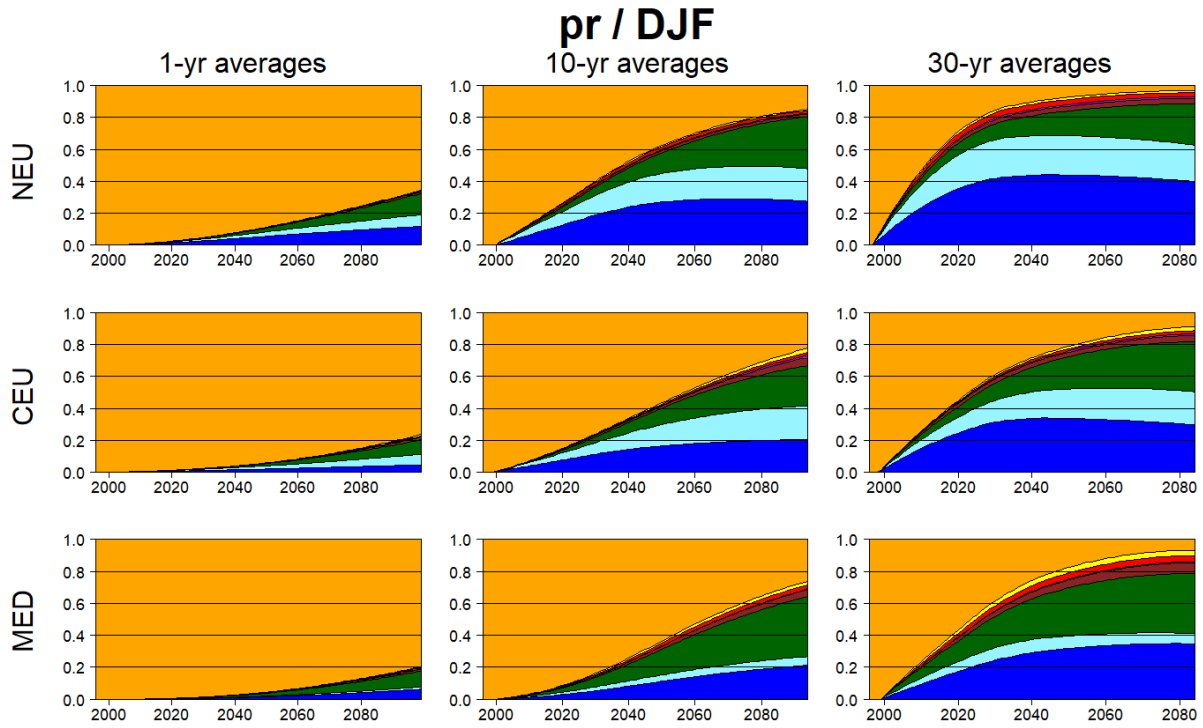

Figure 2: Fraction of total variance for relative changes of total precipitation in winter (DJF), as a function of time, for different temporal aggregation scales: 1-year (left plots), 10-year (middle plots) and 30-year averages (right plots), for the 3 SREX regions.

The revised manuscript also acknowledges that some limitations of the internal variability estimates (subsection 8.4): "*Refined estimates of internal variability could also be considered to take into account 1/ different internal variability by GCM, 2/ different internal variability as a function of time, 3/ the potential autocorrelation present in the deviations from the climate change response, as a result of the statistical preprocessing of the climate projections (in our case, 30-year moving averages).*"

RC1#7: It will also be helpful to provide more introduction and discussions of QUALYPSO so that readers can understand how this method can provide a balanced estimate without reading the reference paper and the codes.

Thank you for this comment. The difficult understanding of the QUALYPSO method was also raised by the reviewer RC2 (comment RC2#3). The revised version of the manuscript now presents a longer description of the method, and is illustrated with the additional figure 1 which includes examples of the different quantities estimated for the different steps of QUALYPSO.

Concerning how this method can provide a balanced estimate, we now recall that "*Missing climate projections are part of the inference and the posterior distributions of all unknown quantities (grand mean, main effects and missing climate projections) are sampled sequentially using the Gibbs algorithm. In particular, following the so-called data augmentation method, missing climate projections are simulated directly from Eq. 1, assuming that they follow the same ANOVA decomposition than available chains.*" Fig. 1c illustrates this aspect with an ensemble of available and missing (generated) climate change responses.

RC2: Anonymous Referee #2

RC2#1: The submitted paper holds a thorough analysis of seasonal mean changes in European temperature and precipitation for the mid-century and end-of-century. It utilises the CMIP5-based EUROCORDEX multi-model GCM-RCM experiments, and a specific method to estimate the so-called 'balanced' climate change response and 'balanced' uncertainty. This uncertainty is then attributed to GCM, RCM, RCP (or interactions between those, or internal variability.

This is an important exercise, and hence I am of the opinion that the paper contributes usefully to the existing scientific literature. I recommend acceptance of this manuscript for publication, though would recommend major revisions to the text before doing so. I will try to explain my discomfort with the text in its current form below.

We thank the reviewer for this positive feedback and for these numerous constructive comments that certainly helped to improve the manuscript.

RC2#2: As noted, the results of the study are important, and go beyond existing work and published analyses of EUROCORDEX as far as I know. However, given the conclusions and the recommendations made there, I would like to see much more comparisons between the QUALYPSO method and the 'normal' approach. This to inform the reader of the gains made and the errors that could otherwise be introduced. I also wonder if a sensitivity analysis, e.g. by splitting the ensemble in two smaller ensembles, computing two times the balanced response and the normal response, and showing (rather than telling and trusting) that this is indeed a robust method.

We fully understand this concern and this type of comparison was actually considered in a first version of this paper. An additional comparison has thus been carried out and is now included in this paper. An important aspect of this sensitivity analysis is to keep the same set of RCMs/GCMs since, as shown in the paper, the overall contribution of GCM/RCM uncertainty is greatly dependent of a few RCMs/GCMs, regardless of the approach for partitioning the uncertainties.

To perform this comparison, we rely on a complete synthetic MME composed of 9 GCMs x 13 RCMs x 3 RCPs = 351 climate change responses generated using ANOVA effects and residual variability estimated with the original MME. We then subsample randomly 1,000 different MMEs of 87 chains among this complete synthetic MME of 351 chains, with at least one chain for each of the 9 GCMs, 13 RCMs and 3 RCPs. Figures 3 and 4 below show the mean change estimates (BM and M) obtained for temperature and precipitation, respectively, for the different RCP scenarios, SREX regions and seasons. Clearly, the variability between the estimates M obtained with direct averages of chains available for each scenario is larger than when QUALYPSO is applied (BM estimates). For temperature, mean estimates obtained with QUALYPSO are particularly stable.

Note that many other experiments could have been done. Here, we ignore internal variability, in order to minimize the computational burden (estimation of the climate change response). This random subsampling is obviously the simplest way to perform this experiment. More realistic subsampling strategies reflecting the overrepresentation of the scenario RCP8.5 or of some RCMs/GCMs have been tested and do not change the conclusions. Similarly, tests with smaller subsampled MMEs lead obviously to larger variabilities between the different samples which are relevant for other CORDEX domains often counting fewer simulations.

These results have been included in subsection 8.1 "*Unbalanced MMEs and robustness of QUALYPSO estimates*" of the revised manuscript.

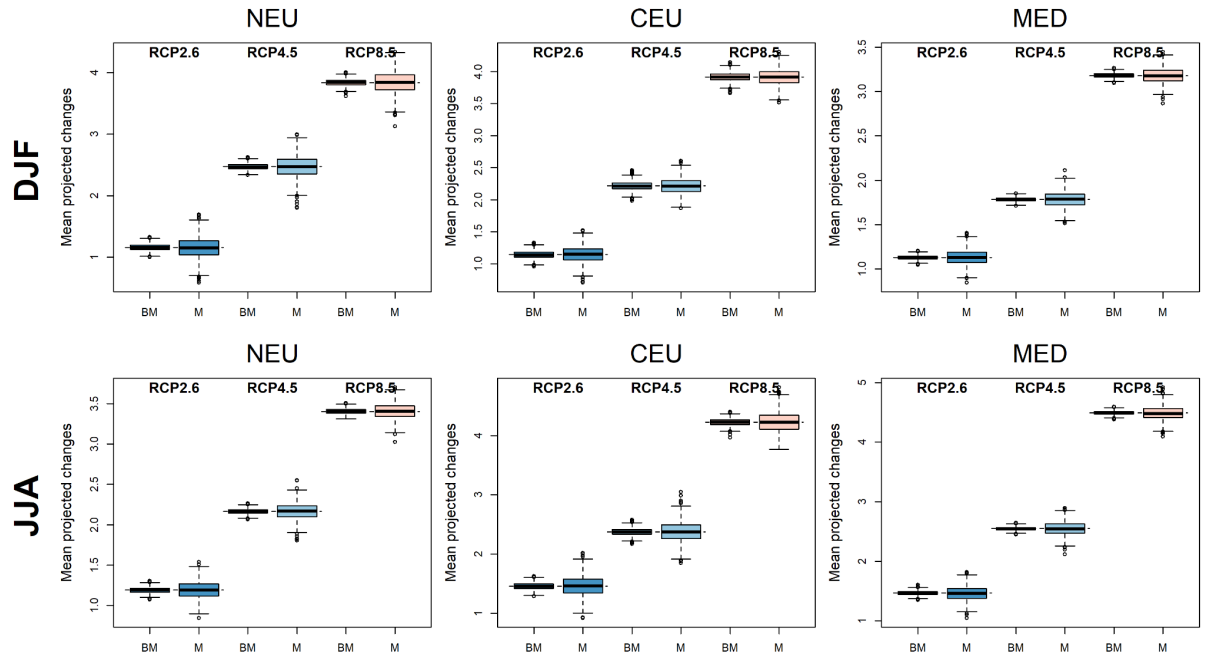

Figure 3: Comparison of mean projected changes estimates for temperature using QUALYPSO (BM) and direct averages (M) of a synthetic MME for each RCP scenario, SREX region, and season. A complete synthetic MME composed of 9 GCMs x 13 RCMs x 3 RCPs = 351 chains is generated using ANOVA effects and residual variability estimated with the original MME. The boxplots show mean change estimates based on 1,000 random subsampling of 87 chains among the complete synthetic MME of 351 chains, with at least one chain for each of the 9 GCMs, 13 RCMs and 3 RCPs. Dashed horizontal lines indicate the corresponding averages obtained from the complete MME.

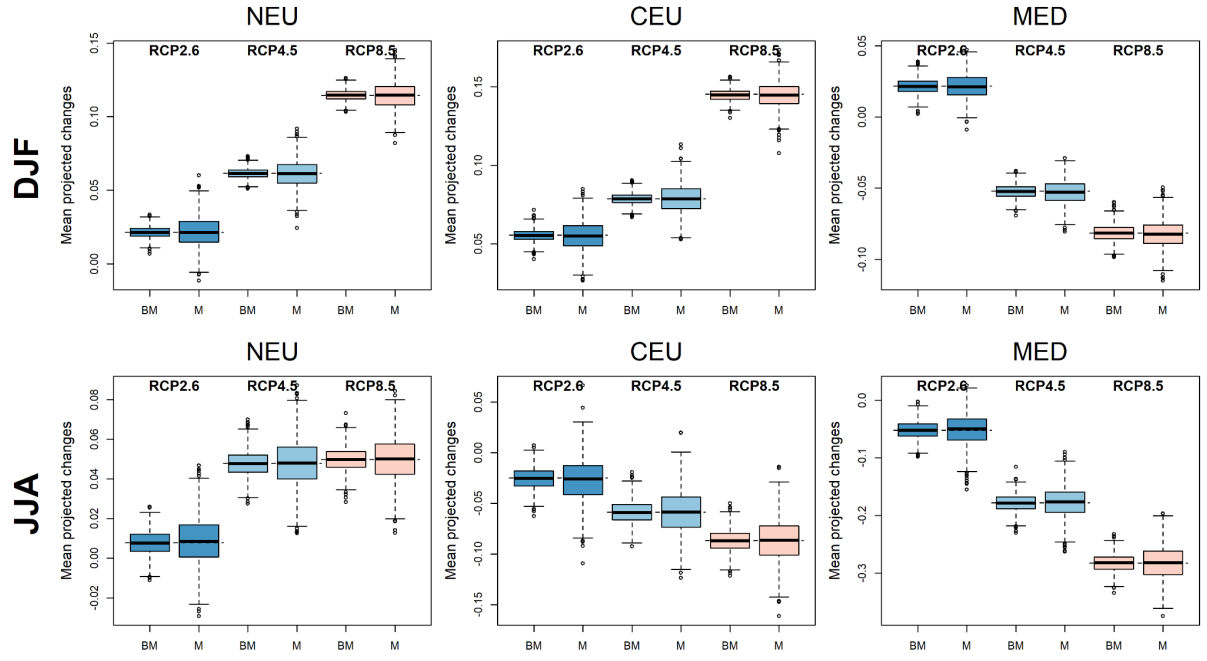

Figure 4: Comparison of mean projected changes estimates for precipitation using QUALYPSO (BM) and direct averages (M) of a synthetic MME for each RCP scenario, SREX region, and season. A complete synthetic MME composed of 9 GCMs x 13 RCMs x 3 RCPs = 351 chains is generated using ANOVA effects and residual variability estimated with the original MME. The boxplots show mean change estimates based on 1,000 random subsampling of 87 chains among the complete synthetic MME of 351 chains. Dashed horizontal lines indicate the corresponding averages obtained from the complete MME.

RC2#3: In its present form, the paper does not excite the reader, or invite thorough reading. The method section is confusing, and does not offer a clear explanation of QUALYPSO to the reader:. I suggest a more understandable and intuitive explanation is added, maybe with some drawn schematics that show how differences between model runs would be quantified in the balanced outcomes and uncertainty quantifications.

We appreciate this comment and we have tried to give a less technical presentation of QUALYPSO in the revised version of this manuscript, as this point was also raised in the comment RC1#7. The new figure 1 now illustrates the different steps of the method.

RC2#4: The results sections feel like a dump of many figures and tables, with lots of text that tell us what can be seen in the figures. It would be more useful to explain sources of differences (e.g. climate sensitivity of GCMs is only noted once in line 332!).

In this study, it is true that we first aim at describing obtained projected mean changes and related uncertainties. However, different attempts are made toward their interpretation according to our knowledge and the literature. In particular, Section 7 dedicated to the individual contributions of each climate model provides many insights that can explain the behaviour of the different climate models. Some simple interpretations of the mean changes, which were lacking in the original manuscript, have been added in Section 5 of the revised manuscript, noting however that explaining the climate change pattern in Europe is not the main scope of our study and has been done widely before.

We fully agree with the reviewer that more efforts could be put into tracking the sources of the model differences as this is undoubtedly an interesting and relevant topic. However, it likely requires a dedicated study for each model family, each variable, geographical region, or season which is out of the scope of the current study. In particular, to be convincing, looking for the origin of the differences between models probably requires new extra-simulations with sensitivity tests. We, therefore, decided to rely on the published literature known by the authors to give some hypotheses. We acknowledge that this approach remains limited but we hope that further studies, starting from our assessment, may explore those differences in depth. In particular, defining the equivalent of the GCM ECS but for the RCMs would be very relevant but does not fit in the current study.

RC2#5: The final paragraph of the manuscript is the strongest paragraph of the text by far. I don't think these conclusions currently follow from your written text, but can see

how they would relate. Rewriting parts of the manuscript, and adding comparisons/sensitivity experiments I believe would allow these conclusions to be made, and indeed would then form a very welcome addition to the literature.

Thank you for your encouragement. We hope that the additional results better support these conclusions. In particular, the synthetic analysis discussed in comment RC2#2 now illustrates the importance of an uncertainty assessment method that is less sensitive to the subsampling of the possible runs. In addition, the conclusion has been modified to avoid conclusions that are not based on the materials shown in the paper. In particular, we have replaced "*This work urges the community*" with "*Following the results shown in this work, we advise the community*" which seems more moderate, as these advices are not strongly related to some particular results.

RC2#6: Section 5 - The balanced mean response is very clear, in line with previous analyses/IPCC and well explained. The explanation of scenario-excluded uncertainty however is rather vague. This also goes back to my questions on the methodology, is this uncertainty equal for each scenario? By design of the method, or also in reality? I would have expected differences in, for example, the influence of internal variability and model-interactions between RCPs.

This is correct, the scenario-excluded uncertainty is equal for each scenario by design of the method. The main reason is that the main effects of each GCM (resp. RCM) are expected to be better estimated when the data from all scenarios are considered together. It could be considered to have a different GCM uncertainty for each scenario by adding a GCM/RCP interaction effect in the ANOVA, but would necessitate estimating 9 x 3 = 27 more terms. In this study, we provide very crude estimates of some of these interactions to decompose the residual variability, but these estimates are not considered to be robust enough to provide different uncertainties for each scenario. Concerning internal variability, it is mostly a matter of simplification. Indeed, we first estimate a different internal variability for each run of the MME. It is then combined by estimating a single internal variability for the whole MME, but we could easily consider a different internal variability for each RCP, or for each GCM. These limitations and possible extensions of the current approach have been added in the section "Discussion":

"QUALYPSO is not free of limitations. For example, the different inferred quantities are not scenario-dependent. Internal variability, the scenario-excluded uncertainty BU, GCM and RCM uncertainties, are thus considered to be identical for the three emission scenarios, whereas some differences could be expected due to the different responses of the GCMs for the different scenarios. In future studies, a different GCM uncertainty for each scenario could be considered. However, for the GCM uncertainty, GCM/RCP interaction effects would have to be estimated in the ANOVA. While very crude estimates of some of these interactions are provided in this study, these estimates probably lack of precision since they rely on a few simulation chains only. Refined estimates of internal variability could also be considered to take into account 1/ different internal variability by scenario and/or GCM, 2/ different internal variability as a

function of time, 3/ the potential autocorrelation present in the deviations from the climate change response, as a result of the statistical preprocessing of the climate projections (in our case, 30-year moving averages).”

RC2#7: In figures 2-5 I find myself mostly looking at the last column. This is where you split and attribute the uncertainty to the different factors. I wonder if having, for all these figures, all other panels are worth adding. I suggest the authors carefully consider this (I’m not advising any direction, just encouraging thinking this through), fewer panels would allow them to be bigger and make them easier to read and interpret.

Thank you for this suggestion. We have followed your advice and merged the last columns into one single figure. The other columns 1:4 have been put in the SM, in separate figures. As a consequence, this section has been largely rearranged.

RC2#8: Fig 2: can you remove the black lines around the colour shading? It prevents us from seeing the colours of narrow shading/small contributors.

Ok, this has been done.

RC2#9: Fig S1a - modify the colour scale so it shows some information please.

Ok, this has been modified in the revised version.

CC2: Comments by Richard Rosen

The context for articles like this one is that the IPCC reports (and many other reports and journal articles) claim that certain climate parameters have a fairly precise probability of occuring or being exceeded by X%. For example, the IPCC WGI reports make statements like the probability of a certain scenario exceeding a 2.0 degree C global temperature increase by 2100 is 66% or greater. But as acklowedged in the comments to this article, this precentage does NOT represent a real-world chance of occurring as the text usually makes it seem, but it merely represents where a given result resides with respect to a range of model results. Thus, such a percentage like 66% implicitly says more about the range of model structures and input assumptions, than it says something truthful about the real world's climate. However, most readers of these reports and most policy makers do not understand this key point.

The problem I find with this article is that this basic conceptual difference is not clearly spelled out and is not clearly incorporated into the analysis at each relevant point. But it must keep reminding the reader when the uncertainties being discussed apply to the real world physical processes of climate change, or to the models individually, or to the range of model results collectively. Similarly, the uncertainties associated with each of these three levels must be clearly distinguished throughout the text. Yet, this will be very difficult to accomplish properly, and it is almost never accomplished in similar articles in

the literature, whether dealing with regional or global models. Conceptually, it does not matter what the scope of the model is. But, in general, uncertainties usually increase in percentage terms when smaller areas of the earth's surface are modeled in comparison to when global models are run. This fact is due to flucuating and difficult to model weather patterns at the regional level.

As far as I know, the "bottom line" of attempts like this article at uncertainty analysis is that it is fairly hopeless to succeed at results that have much physical meaning or policy making value. Consider the results for global climate sensitivity. Roughly the results for many years have been 3.0 degrees C plus or minus 1.5 degrees C. This range is primarily due just to the differences in model results within an ensemble, without consideration of the other two kinds of uncertainty noted above. But even considering this one kind of uncertainty, the size of the uncertainty band is comparable to the size of the effect being forecast, i.e. an uncertainty range of 3.0 degrees C for a likely effect of 3.0 degrees C. Thus consideration of the other two kinds of uncertainty will only increase the total range of uncertainty relative to the median projection of about 3.0 degrees C. So what is the point of trying to perform a total uncertainty analysis on a firm scientific basis. I suggest that the result is basically unknowable. Of course, if we just look at the history of global climate change from about 1978 to today (2021) we find a fairly steady increase at about 0.15-0.20 degrees C per decade in actual fact, which will give policy makers a far more likely range of uncertainty in global climate change if projected linearly for at least the next few decades (at least to 2050), which is the most important time period if climate change is going to be effectively mitigated. The results of a more sophisticated and comprehensive uncertainty analysis taking all three conceptual levels of uncertainty into account is not likely to be more useful for policy makers (or anyone else) at this point in time.

Thus, I do not believe that this paper rises to the level of useful publishable material unless the ensemble model results can be connected to the real history of the world in a way that reduces the uncertainty range and does not increase it. But I think this is impossible given the complex system involved.

We partly agree with these comments by Richard Rosen. Indeed model-based articles such as our own study are dedicated to the analysis of model ensemble results and not directly to the real future world. This is the case for all model results in science as models are built to be a simplified version (approximation) of the real world whatever the scientific domain. Therefore simulations of the future made with those models are not deterministic forecasts of the real world but projections or scenarios that can under some conditions inform decisions (but not always). In particular, we agree that the question of model simulations being representative or not or partly representative of the real world or representative at certain spatial and temporal scales is indeed very relevant.

To our knowledge, IPCC report statements (or other expert-judgment-based reports) are of a different kind. Indeed, the statements and uncertainty quantification of those reports are based on expert judgments, including model ensemble results as only part of the

input information to establish the judgment. So we prefer to clearly distinguish between model-based studies such as our article and climate expert reports.

This said we agree that, in our study, given uncertainty ranges inform about the model structure and modelling assumption. If the models used are well designed and built (this is what we trust) and if the RCP scenarios are representative of the future GHG trajectory, then we can hope that the derived climate change values and associated uncertainty range are informative about plausible future climates.

Concerning the last point "However, most readers of these reports and most policy makers do not understand this key point", it is difficult to say as we are not specialists of social sciences but we agree that explaining the meaning of model-based climate change scenarios to potential users of the information (incl. policy makers) needs more effort from the whole climate community. Note however that this study, published here in a specialized journal, is mostly at the destination of the climate modellers community and expert climate model users community and not the policy makers.

However, we agree that the estimated mean climate change response and associated uncertainties cannot be considered as predictions of the future climate properties (see also our response to comment RC1#5) and we do not use the terms "predictions" or "predict" in the manuscript.

References

Benestad, Rasmus, Kajsa Parding, Andreas Dobler, and Abdelkader Mezghani. 2017. "A Strategy to Effectively Make Use of Large Volumes of Climate Data for Climate Change Adaptation." Climate Services 6 (April): 48–54. https://doi.org/10.1016/j.cliser.2017.06.013.

Dosio, Alessandro. 2016. "Projections of Climate Change Indices of Temperature and Precipitation from an Ensemble of Bias-Adjusted High-Resolution EURO-CORDEX Regional Climate Models." *Journal of Geophysical Research: Atmospheres* 121 (10): 5488–5511. https://doi.org/10.1002/2015JD024411.

Gutiérrez, J. M., D. Maraun, M. Widmann, R. Huth, E. Hertig, R. Benestad, O. Roessler, et al. 2019. "An Intercomparison of a Large Ensemble of Statistical Downscaling Methods over Europe: Results from the VALUE Perfect Predictor Cross-Validation Experiment." *International Journal of Climatology* 39 (9): 3750–85. https://doi.org/10.1002/joc.5462.

Hingray, B., Evin, G., Blanchet, J., Eckert, N., Morin, S., and Verfaillie, D.: Partitioning uncertainty components of an incomplete ensemble of climate projections using data augmentation, EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-21864, https://doi.org/10.5194/egusphere-egu2020-21864, 2020.

Jacob, Daniela, Claas Teichmann, Stefan Sobolowski, Eleni Katragkou, Ivonne Anders, Michal Belda, Rasmus Benestad, et al. 2020. "Regional Climate Downscaling over Europe : Perspectives from the EURO-CORDEX Community." Regional Environmental Change 20 (2). http://urn.kb.se/resolve?urn=urn:nbn:se:smhi:diva-5677.

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth,B., Coppola, E., Vries, H. d., Harris, G., Hegerl, G. C., Knutti, R., Lenderink,G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi,P., and Undorf, S. (2020). Comparing Methods to Constrain Future Euro-pean Climate Projections Using a Consistent Framework. Journal of Climate,33(20):8671–8692.

Evin, G., Hingray, B., Blanchet, J., Eckert, N., Morin, S., and Verfaillie, D. (2019). Partitioning Uncertainty Components of an Incomplete Ensemble of Climate Projections Using Data Augmentation. Journal of Climate, 32(8):2423–2440.

McSweeney, C. F., R. G. Jones, R. W. Lee, et D. P. Rowell. « Selecting CMIP5 GCMs for Downscaling over Multiple Regions ». *Climate Dynamics* 44, nº 11 (1 juin 2015): 3237-60. https://doi.org/10.1007/s00382-014-2418-8.

Vrac, Mathieu, et Petra Friederichs. « Multivariate—Intervariable, Spatial, and Temporal—Bias Correction ». *Journal of Climate* 28, nº 1 (30 septembre 2014): 218-37. https://doi.org/10.1175/JCLI-D-14-00059.1.