We are grateful for the evaluation of our paper and all the useful comments and suggestions. We have considered all of them and revised the manuscript accordingly. Please find below our responses to specific comments, the revised manuscript is submitted as a supplement to a separate author comment.

The analysis uses only spatially averaged time-series information, unlike earlier work (e.g. Knutti 2013, Sanderson 2015) which primarily use spatial bias correlation to assess similarity. By not using spatial information, it seems like the authors are throwing away a lot of potentially useful information. This is not a showstopper - but the authors should acknowledge that by using both spatial and temporal information, more meaningful results could probably be obtained

It is certainly true that the evaluation of spatial simulated fields is important. But in current study we have chosen to concentrate on temporal behaviour of the time series averaged over the large European regions. Comparison of spatial fields from RCMs and GCMs is complicated, mainly by large differences in spatial resolution and also by differences in effective spatial resolution (which depends on numerical methods incorporated in the models). We have not figured out how the spatial information could be incorporated in our current setting of the methodology. Spatial fields from GCMs are much smoother than RCMs, and therefore if we convert the fields into functions, the results will be very different in nature. By smoothing (regridding) the RCM fields to GCM-like coarse resolution would result in throwing away a lot of information. But it is probably a good topic for another possible application of our methodology framework, to apply it for evaluation of spatial simulated fields, but for an ensemble consisting of simulations with comparable spatial resolution.

some parameter sensitivity is required - or at least an explanation of why some arbitrary decisions were made. The domain averaging size, for example - a larger averaging area for precipitation might result in a less noisy field in which model similarities are more accurately identified. Similarly, the averaging period and the parameters of the spline expansion - how sensitive is the method to these choices?

The domains used in our study are the quite large “PRUDENCE” regions very often used for analysis of RCM outputs over Europe. The results for smaller regions would probably be more influenced by internal variability and differences between RCMs connected to smaller scale processes and orography representation.
Regarding the averaging period, we have not evaluated the sensitivity of the method. The choice of the length of the period is not basically an arbitrary choice, but it originates in the fact, that we intended to work with long-term means as the main characteristics of climate. And 30-year period is as far as our knowledge the most common period length used in climatology.
Regarding the parameters of the spline expansion, we have analysed the sensitivity of d_0 and d_1 distances on the amount of smoothing of the underlying curves. The results for an arbitrarily chosen example are shown in Supplement2 and commented on in the end of Section 3.1. The results do not strongly depend on the smoothing. The dependence is slightly stronger for d_1, but even for that the structure of the distances is quite stable for the whole ensemble.

what is the expected noise from climate variability, and can this be quantified more accurately? Can the authors use initial condition ensemble members to identify the expected intermodel distance which arises from climate variability alone?

The influence of internal variability on RCM simulation is difficult to be evaluated, as simulations with perturbed initial conditions are not available (as far as our knowledge). Earlier findings (Déqué et al., 2007, Déqué et al., 2012, Hawkins and Sutton, 2009, 2010) suggest that the influence of internal variablity on the overall uncertainty of simulated air temperature and precipitation changes is expected to be rather low. To investigate the issue we compared the results for the ensemble used in

our study with a mini-ensemble consisting of 5 simulations of CNRM-CM5 GCM with perturbed intial conditions (runs denoted as r1i1p1, r10i1p1, r2i1p1,r4i1p1, r6i1p1). We chose this GCM to maximize the number of RCMs driven by it and the extent of resulting mini-ensemble. The figures are available in Supplement3 and the results are commented on in the last section of the revised paper:
„As explained in the Introduction, the spread of multi-model ensembles is considered as an estimate of structural model uncertainty. For analysis of the influence of internal variability on the overall uncertainty, simulations with perturbed initial conditions can be used. Unlike GCMs, for RCMs these are not generally available. In Supplement3 a suite of figures showing FDA similarities between 5 simulations of CNRM GCM with perturbed initial conditions is provided. The aim of these figures is to illustrate the range of uncertainty stemming from internal variability. We chose CNRM GCM to maximize the number of RCMs driven by this GCM and the number of mini-ensemble members. The figures suggest that for air temperature changes the spread of the CNRM mini-ensemble covers almost a half of the multi-model ensemble spread (Fig. S3.1). In case of precipitation, the portion of the spread is smaller (Fig. S3.2). The d_0 and $d_1$distances between the members of CNRM mini-ensemble are shown in Fig. S3.3 – S3.6. To enable the comparison with the distances for the multi-model ensemble, their values before normalization are provided in Fig. S3.7-S3.10. For air temperature, the maximum inter-model distances are almost twice as large as the inter-simulation distances within the CNRM mini-ensemble (compare Fig. S3.3, S3.4 and S3.7, S3.8). In case of precipitation, the $d_0$distances between the simulations with perturbed initial conditions are very small in comparison to inter-model distances (Fig. S3.5 and S3.9). However, for $d_1$distances the difference is not so struggling (Fig. S3.6 and S3.10). The fact that the range of uncertainty connected to internal variability is relatively larger (in comparison to structural uncertainty) for air temperature than for precipitation probably points to larger overall structural uncertainty in case of precipitation than air temperature, i.e. the inter-model differences in simulation of processes connected to precipitation changes are larger than in case of air temperature changes. However, we have to keep in mind that presented results rely only on a limited number of simulations from one GCM."

Déqué, M., Rowell, D.P., Lüthi, D., Giorgi, F., Christensen, J.H. et al., 2007. An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. Climatic Change, 81, Supplement 1, 31–52.

Déqué, M., Somot, S., Sanchez-Gomez, E., Goodess, C. M., Jacob, D., Lenderink, G., Christensen, O. B. (2012): The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability Climate Dynamics, 2012, 38, 951-964

Hawkins, E., Sutton, R., 2009. The potential to narrow uncertainty in regional climate predictions. Bulletin of the American Meteorological Society. DOI: 10.1175/2009BAMS2607.1

Hawkins, E., Sutton, R., 2010: The potential to narrow uncertainty in projections of regional precipitation change. Climate dynamics. DOI: 10.1007/s00382-010-0810-6

the graph plots are nice - but there are precedents in the literature for presenting model similarities in 2D space, which should probably be cited here (Sanderson 2015).

We have added a comment to the last section:

"Unlike similar approach of multidimensional scaling used in Sanderson et al. (2015), which also results in 2-dimensional visualization of inter-model distances, the layout graphs do not require defining any data node as a central (reference) point of the whole ensemble."

I feel slightly more could be made of the discussion of parent GCMs and embedded RCMs. Figure 4 suggests that the parent GCMs dominatethe inter-model distances for both d0 and d1 for temperature, but perhaps not for precipitation where there is clear structure from RCM pairs. This is perhaps one of the more interesting results from the paper - and the authors should make more of it. Why is this the case, what are the mechanisms? What recommendations would the authors give for end-users of CORDEX given this finding?

The mechanisms for different results for DJF tas over BI and JJA pr over EA are commented on in the paper (Section 5) :
"It is clearly seen that when large-scale phenomena are responsible for output, as in case of temperature changes over BI region, RCMs tend to be very close to driving GCM, and different GCMs are apart from each other (Figs. 1 and 7). On the contrary, when smaller scale processes are more in play, such as in case of JJA precipitation changes over EA, the results are more influenced by RCMs (Figs. 2 and 8). This does not automatically imply any real added value in the sense of more realistic simulation. Rather, it points to differences in implementation of the local processes in different RCMs. In our case, different parameterization schemes employed to simulate convection, microphysical processes in clouds and surface processes including soil moisture are possible candidates."

Our results are not really representative for air temperature and precipitation over the whole European domain and for all seasons, but it illustrates that there are large differences between individual cases. Therefore, a recommendation for end-users is that an analysis of GCM-RCM interactions and a thorough choice of representative simulations (if it is not possible to use the whole multi-model ensemble) for impact studies is necessary. Our paper offers a tool for such analysis. We added a comment on this into the last sections:

"The results of presented case study for two basic climatic variables over two European regions show that the structure of the multi-model ensemble and the GCM-RCM interactions can differ substantially in individual cases. Therefore, before the RCM outputs are used in any applied research (e.g. studies on impacts of projected future climate changes) an analysis of GCM-RCM interactions and a thorough choice of RCMs to be used is necessary. Present paper offers a convenient tool for this purpose."