We thank both anonymous reviewers for their helpful and constructive comments. Please find below our replies (author comments, AC) to each individual reviewer comment (RC), as well as a listing of the actual changes to the manuscript. We hope the suggested changes to the manuscript are considered appropriate, we'd be looking forward to submit a revised version of the manuscript.

Comments of reviewer #1

***RC1**. This paper presents an evaluation of the EURO-CORDEX ERA-Interim driven simulations. The analysis focuses on eight large sub-regions of Europe and documents model performance for temperature and precipitation. The study also places these results into context with the previously performed ENSEMBLES project. The study is presented well and is likely to become a standard reference for anyone working in EURO-CORDEX. In many ways it is a nice summary of the RCMs performance and is certainly a worthwhile contribution to the field. I recommend publication subject to the technical corrections below.*

AC1. We are pleased about this positive assessment by reviewer #1. All suggested technical corrections will be included in the revised version of the manuscript.

Related changes to the manuscript:

- All suggested corrections were included.

Comments of reviewer #2

***RC2.** General comments: This paper shows us the preliminary results of EURO-CORDEX project, which is the revised version of the pioneering project, EU-ENSEMBLES, by using higher resolution regional climate models. All the results written in the paper does not conflict with the results of all the existent research of dynamical down-scaling. It looks like a well written report of the experiment, but it is hard for us to find scientifically new thing in the paper.*

AC2. As outlined in Chapter 1, one of the primary aims of our study is to demonstrate and document the status of the new EURO-CORDEX ensemble applying the most recent model generation. This is done based on monthly and seasonal mean values, averaged over larger sub-domains over the continent and applying eight individual performance metrics. The results are compared against those obtained for the previous (and widely used) ENSEMBLES experiments. As such, the analysis provides a comprehensive overview and update on RCM performance over Europe but leaves individual aspects to subsequent studies (some of which are already published, such as Vautard et al. 2013, or have been submitted in meantime). The evaluation methods themselves are not intended to be "scientifically new", they are rather established approaches in model evaluation so as to enable intercomparison against previous studies. The work itself is of course original. It provides benchmarks for future evaluation efforts. The fact that we focus on documenting the performance of a new model ensemble using established performance measures was exactly the reason why we did chose GMD as journal. Two of the four manuscript types considered for publication in GMD are (1) Papers describing new standard experiments for assessing model performance, or novel ways of comparing model results with observational data and (2) Model intercomparison descriptions, including experimental details and project protocols (see http://www.geoscientific-model-development.net/home.html). We believe that our study falls indeed into these two areas. The related manuscript type (see

http://www.geoscientific-model-development.net/submission/manuscript_types.html) would be "Model Assessment Methods papers" and "Model Experiment Description papers".

At this point, we'd also like to express our disagreement with the reviewer's note that only "preliminary results" are presented. This is not the case, as the EURO-CORDEX simulations (both the evaluation runs investigated in the present work and climate scenarios) have already been or are currently being published on the ESGF archive. A proper evaluation of the applied RCMs is an important information for end users. As new model runs according to the EURO-CORDEX protocol and applying further RCMs are likely to become available in the future, our analysis is certainly not complete in each and every aspect and might have to be updated at a later point in time. However, it includes those experiments available at the time of paper submission, representing the majority of the final set of models. To better clarify this point we propose to explicitly mention possible future extensions of the EURO-CORDEX model set in the introductory Chapter 1.

Related changes to the manuscript:

- page 8, lines 29-31: Added comment wrt. possible future extensions of the EURO-CORDEX ensemble.

RC3. *In the paper, they validate the results of many models nesting to ERA-Interim, but there is no explanation of the characteristics of ERA-Interim itself. We can easily find that the parent GCM had great influence on the calculation results of RCM. Thus, we would like to know the systematic bias appeared in ERA-Interim, before discussing the results of down-scaling results. For the same reason, we also would like to know the character of ERA40, compared to ERA-Interim, because it would affect the difference between 0.22 degree grid RCM used in EU-ENSEMBLES and 0.11– 0.44 degree grid RCM used in this paper.*

AC3. We certainly agree that the choice of the boundary forcing can significantly affect the RCM performance, especially when nesting a specific RCM into different GCMs. However, in case of re-analysis forcing (as in our study) the effects of different boundary data (i.e., the choice of different re-analysis products) are expected to be much smaller as all re-analyses are largely confined by the assimilation of observational data. See for instance the work of Lucas-Picher et al. (2013) that, among other aspects, investigates the influence of the driving re-analysis (ERA-Interim vs. ERA40) on RCM results over North America. Still, we certainly recognize that the quality of ERA-Interim is very probably superior to ERA40 in many aspects which is partly connected to a more sophisticated assimilation scheme (4D-Var vs. 3D-Var; Dee et al. 2011). As for the comparison of the EURO-CORDEX ensemble against the ENSEMBLES experiments presented in the manuscript, there are further differences in addition to the different driving re-analysis (different analysis period, different ensemble size) which cannot be avoided and which are openly mentioned in the first paragraph of Section 4.6. We here propose to include a brief discussion of the influence of the driving re-analysis also in Section 2.1 of the manuscript in order to highlight this issue more prominently, and to also cite the study of Lucas-Picher et al. (2013) at that point.

Concerning the quality of the driving ERA-Interim reanalysis itself and the influence of potential biases in large-scale fields on the downscaling there is no straightforward way to proceed in our opinion. An evaluation of ERA-Interim's 2m temperature and precipitation over the analysis domains would not serve this purpose, as the RCMs actually do not use these fields as input, but prognostic atmospheric 3D quantities within a sponge zone along the lateral boundaries of each individual RCM domain. These quantities would have to be evaluated in ERA-Interim over the respective sponge zones, but reference data to evaluate against do not exist. The only possibility would be to compare different re-analysis products against each

other, i.e., to analyze re-analysis uncertainty. Such an assessment has actually been presented by Brands et al. (2013), showing that re-analysis uncertainty (in their case ERA-Interim vs. JRA-25) is negligible along the sponge zone of the EURO-CORDEX domain, but can be considerable for the African CORDEX domain. Their results for Europe provide additional confidence in the good quality of the driving ERA-interim re-analysis for EURO-CORDEX applications. We propose to briefly discuss this issue in Section 2.1 of a revised manuscript, explicitly citing the work of Brands et al. (2013).

Lucas-Picher P, Somot S, Déqué M, Decharme B, Alias A, 2013 : Evaluation of the regional climate model ALADIN to simulate the climate over North America in the CORDEX framework. Climate Dynamics 41: 1117-1137.

Brands S, Herrera S, Fernández J, Gutiérrez JM, 2013: How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? A performance comparison for the downscaling community. Climate Dynamics 41: 803-817.

Related changes to the manuscript:

- page 8, lines 11-14: Added comment on the quality of the driving ERA-Interim re-analysis.
- page 9, lines 13-16: Added comment on the influence of the different driving re-analysis on RCM results in EURO-CORDEX vs. ENSEMBLES.

RC4. *In this paper, they validate the data in monthly to seasonal time scale and 8 sub-domains in Europe, and could not find the advantage of using high resolution RCMs. Our impression is that the validation time-scale and space-scale is too coarse to find the advantage of high resolution models. As written in Kanamitsu and DeHaan (2011), effect of the higher resolution would appear in a specific region. Thus we should adopt a metric which could find such localized effect.*

AC4. Yes, the reviewer is perfectly right. The validation time and space scales are too coarse to demonstrate an added value of resolution. However, as clearly outlined in the introduction it is not our intention to demonstrate an added value of the higher resolution experiments but to analyze the high and the coarse resolution ensemble on larger scales (European sub domains) which are well represented by both resolutions. The added value of the higher resolved simulations can be identified, but is demonstrated in a number of follow-up studies by the EURO-CORDEX community that are submitted or in preparation (e.g. analysis of precipitation extreme indices). A drawback of these studies is that they do not cover the entire European continent as observational reference data at the required resolution are not available on a European scale. Our reference data, on the other hand, is available for the entire continent but at a nominal resolution of 25 km, i.e., coarser than the EUR-11 ensemble which unfortunately requires an aggregation of the 12 km experiments (as discussed in Chapter 3.3) in order to ensure a consistent comparison. For the revised version of the manuscript we propose to better highlight the scope of the work in Chapter 1 and, furthermore, to include the suggested reference Kanamitsu and DeHaan (2012) as a further example for the added value of high resolution experiments (in addition to Prein et al. 2013a and 2013b, Bauer et al. 2011 and Warrach-Sagi et al. 2013).

Related changes to the manuscript:

- page 7, lines 12-16: Revised explanation of the scope of the present study wrt. added value analysis.
- page 25, lines 5-6: Added reference to Kanamitsu and DeHaan, 2001.

RC5. *In this paper, they avoid the difference of resolution in both observed data and model data, by smoothing the higher resolution data. But could we really compare such different comparison data in the same Taylor-diagram?*

AC5. We do not see any problem for those Taylor diagrams that are addressing the temporal variability (Figs. 11 and 12 of the main manuscript). In there, only averages over the 8 analysis domains are analyzed. For all resolutions (EUR-11, EUR-44 and ENS-22) these averages are computed based on a comparatively large number of grid cells. The reviewer is probably concerned about the spatial Taylor diagrams of Figs. 9 and 10 (and Figs. B5 and B6 of the Appendix). Indeed, the sample sizes analyzed (number of grid cells) are different for EUR-11, EUR-44 and ENS-22. But we are not aware of any systematic effect of the sample size on, for instance, the spatial correlation coefficient or the centered RMSE. From that point of view, a comparison of the respective markers for EUR-11, EUR-44 and ENS-22 makes sense. All comparisons are carried out for the same spatial domain(s) and only datasets reflecting a common resolution (models and observations on 12km, models and observations on 25 km, models and observations on 50 km) are compared against each other. Our proposition for the revised manuscript is to explicitly mention the differing sample size in the captions of Figs. 9 and 10 (and Figs. B5 and B6 of the Appendix).

Related changes to the manuscript:

- Figures 9 and B5: Added note on the different ensemble sizes in the figure caption.

RC6. *In this paper, we could not find any rationality in the selection of 8 sub-domains. It is hard to agree that 8 sub-domains are selected only because "following to PRUDENCE". As they introduce, for AL (alpine) region, there are two quite different climate sectors, one the Alpine mountainous region and other the mount foot plane region, which makes it difficult to analyze the result around there.*

AC6. Your concern is certainly right, and we clearly point out the problem of spatial climatic variability WITHIN the individual sub-domains in Chapter 3.1. The choice of the so-called "PRUDENCE regions" is motivated by the fact that these domains, despite the shortcoming mentioned above, have evolved to standard analysis domains for RCM analysis over the Europe continent, and their definition has been used in a large number of publications during the last 5-10 years and is still widely applied. This fact enables consistent comparisons of different studies. Of course, for more regional analyses over, for instance, the Alpine region a more detailed sub-division is required. Still, the 8 PRUDENCE domains sample important aspects of continental-scale climatic variabilities in Europe and we therefore consider them as an appropriate solution for our European-scale evaluation effort. A division into many smaller sub-domains might allow for a more process-based evaluation in some cases, but would certainly deteriorate the readability and clarity of the manuscript. The horizontal bias patterns of Figs. 2, 3 and 4 actually allow for a more detailed analysis at least for the EUR-11 ensemble, which is indeed included in parts of chapter 4.1 (e.g., concerning the topography-related bias pattern of CNRM).

We'd also like to point out that we did not introduce these regions as "following to PRUDENCE" as implied by the reviewer but, in our opinion, properly motivated their use and cited the relevant reference ("These domains have been specified in the frame of the PRUDENCE project (Christensen et al., 2007) and have since then been widely used for RCM evaluation and analysis of climate change signals."). For the revised version of the manuscript we propose to add a few further citations of recent studies that apply the definition of the PRUDENCE analysis domains.

Related changes to the manuscript:

- page 11, lines 9-10: Added further references to recent studies that are using the PRUDENCE sub-domains.

RC7. *In this paper, they validate 2m temperature and precipitation, because they are "two main parameters required by climate impact modelers". However for agricultural impact modelers, they need also humidity and downward short wave radiation data to drive their crop yielding model (Iizumi et al., 2012). For hydrologist, they also need downward short wave radiation to estimate water budget in some water basins. They need to focus on target, before selecting both the parameter and metrics.*

AC7. As mentioned in Chapter 1, near-surface air temperature and precipitation are standard variables that are analyzed by a diverse set of impact studies. As such, the ability of RCMs to reproduce these quantities is a useful information for a wide range of end users. Furthermore, these two parameters are in the focus of most RCM calibraton/tuning efforts. We agree that further parameters are required by individual fields of climate impact research. Also, a process-based model evaluation with the aim to actually improve certain model deficiencies requires a validation of further parameters, especially process-based quantities such as surface-atmosphere fluxes. Our study, however, does not target specific fields of impact research, but should provide an overview on model performance that is of interest to a broader community. Also, European-scale reference data for variables such as the ones mentioned by the reviewer (relative humidity or downward shortwave fluxes) are not available (or to a limited extent only applying radiative transfer models based on satellite observations). We therefore believe that it makes sense to basically restrict our analysis to near-surface air temperature and precipitation. Further quantities will be evaluated in further ongoing works addressing specific issues of model performance on more regional scales (e.g., basin-wide water budgets, Alpine snow cover). We propose to slightly extend the motivation for restricting our analysis to temperature and precipitation in Chapter 1.

Related changes to the manuscript:

- page 7, lines 4-5: Revised motivation for focusing on temperature and precipitation only.

RC8. *They insist on the effect of topography for the added value in RCMs. However, in the discussion, there is nothing said about the envelope mountain. We understand that in some RCM, they adopt envelope mountains and the precipitation pattern look much coarser than the resolution of the model itself (Fig. 1of Ishizaki et al., 2012). They had better comment on that thing in the paper, too.*

AC8. None of the RCMs analyzed in our study uses the concept of envelope topographies. However, some RCMs apply a filter to the surface orography, thereby smoothing the spatial orographic pattern and avoiding strong orographic grid-cell-to-grid-cell gradients. This is one reason why the spatial precipitation bias pattern in Fig. 3 looks smoother than the nominal RCM resolution in some cases. We propose to explicitly list those RCMs which apply an orography filter and to refer to its effects when discussing Fig. 3.

Related changes to the manuscript:

- page 8, lines 16-18: Added information on orography filtering in individual RCMs.
- page 15, lines 8-12: Added discussion on the possible influence of orography filtering on the results.

RC9. *They comment on the under-catchment of the rain gauge in the paper. We understand that it is a severe issue when we validate the precipitation data with the observation. However, the rate of catchment becomes quite different when the precipitation is rain or snow. In snow case, sometimes the catchment is only around 50%. Thus they need to analyze much more carefully if they think it is a serious issues.*

AC9. The issue of precipitation undercatch in the observational reference is indeed severe. As the reviewer correctly points out the actually registered precipitation amount can be less than 50% of true precipitation in case of snowfall (but typically amounts to less than 20%; see the references cited in the manuscript). In the current version of the paper, this issue is highlighted in Chapter 2.2 and referred to during the discussion of the evaluation results. Explicitly accounting for this undercatch in our evaluation study in a quantitative way would certainly be valuable, but there is no straightforward way to do so. The E-OBS reference data is afflicted with this measurement bias, but a corrected version of E-OBS does not exist. It would be far beyond the scope of our paper to correct E-OBS in this respect. The undercatch of a precipitation gauge considerably depends on the specific location of a site, the prevailing wind conditions, humidity, the phase of precipitation and further factors. Correction therefore requires a set of further parameters measured at a specific site. For most sites underlying the E-OBS precipitation grid this information is not available. We are only aware of one single gridded precipitation dataset that is available both in an uncorrected and a corrected version. This dataset (REGNIE) is provided by the German Weather Service but only covers the area of Germany and would, hence, not serve the purpose of a European scale evaluation. Also, the correction method is based on a rough categorization of station characteristics and does not consider the actual wind velocity or precipitation type (snow or rain) of single events. It therefore only provides a rough estimate of the sampling error but is not suitable to investigate the real influence quantitatively. For the moment, the only possibility we see is to prominently mention the issue of precipitation undercatch when discussing the evaluation results. Additionally, we propose to add a horizontal line at +25% to Figs. 6, 8, B2 and B4 and discuss the location of the markers wrt. this line. Assuming an undercatch of 20% of true precipitation (i.e., only 80% registered), a positive model bias of up to 25% versus the biased measurement would still be acceptable (in case of a 25% bias wrt. an uncorrected reference, the “true” bias would actually be zero in that case).

Related changes to the manuscript:

- page 10, lines 19-23: Added explanation wrt. to precipitation undercatch considered in the revised version of the manuscript.
- page 16, lines 27-29: Added discussion on acceptable model biases given a precipitation undercatch of up to 20%.
- page 17, lines 26-28: Added discussion on acceptable model biases given a precipitation undercatch of up to 20%.
- page 26, lines 30-31: Added brief discussion of acceptable model biases in the light of observational undercatch.
- Figures 6, 8, B2, B4: Highlighted the 0 to +25% bias range by a blue shading and added explanation of the meaning in the figure caption.

Further proposed changes to the manuscript

In meantime, a further explanation for the widespread dry bias of the CNRM model became available. The issue is very probably not related to the model physics but to a detail of the technical setup (namely the choice of the nudging coefficient and a relaxation time of 15 min only outside of the common EURO-CORDEX analysis domain). We propose to include this additional information in a revised version of the paper.

Related changes to the manuscript:

- page 23, lines 14-20: Added explanation CNRM's dry summer bias.

Further changes to the manuscript:

- page 25: Removed last sentence of Section 5.1.
- page 31, lines 6-8: Added acknowledgment for the SMHI experiments.