Author response to the review of Anonymous Referee #2

Referee 2, thank you for your very detailed comments. We are delighted to take all of your comments into account to improve the manuscript. Our answers on all questions, suggestions and remarks can be found on the next pages. Firstly, we summarize the major changes we will make to the revised version of the manuscript based on the comments of the different reviewers:

- *We will include an analysis of the annual cycle over the subdomains as defined by the IPCC6 report (Iturbide et al., 2020) which are situated within the CAS-CORDEX domain. The results, both for the RCMs and the gridded datasets, for the mean temperature and precipitation are given in Fig. A1 and A2.*
- *We will approach the differences between the gridded datasets in a different way. The spread between the gridded datasets (Fig. A3) will be used as an estimate of the uncertainty.*
- *We will improve the discussion section by describing which model features can explain the significant biases that were obtained over certain regions.*
- *We will include some additional recently published scientific papers in our revised manuscript e.g. Harris et al. 2020; Wang et al. 2020; Zhu et al. 2020.*

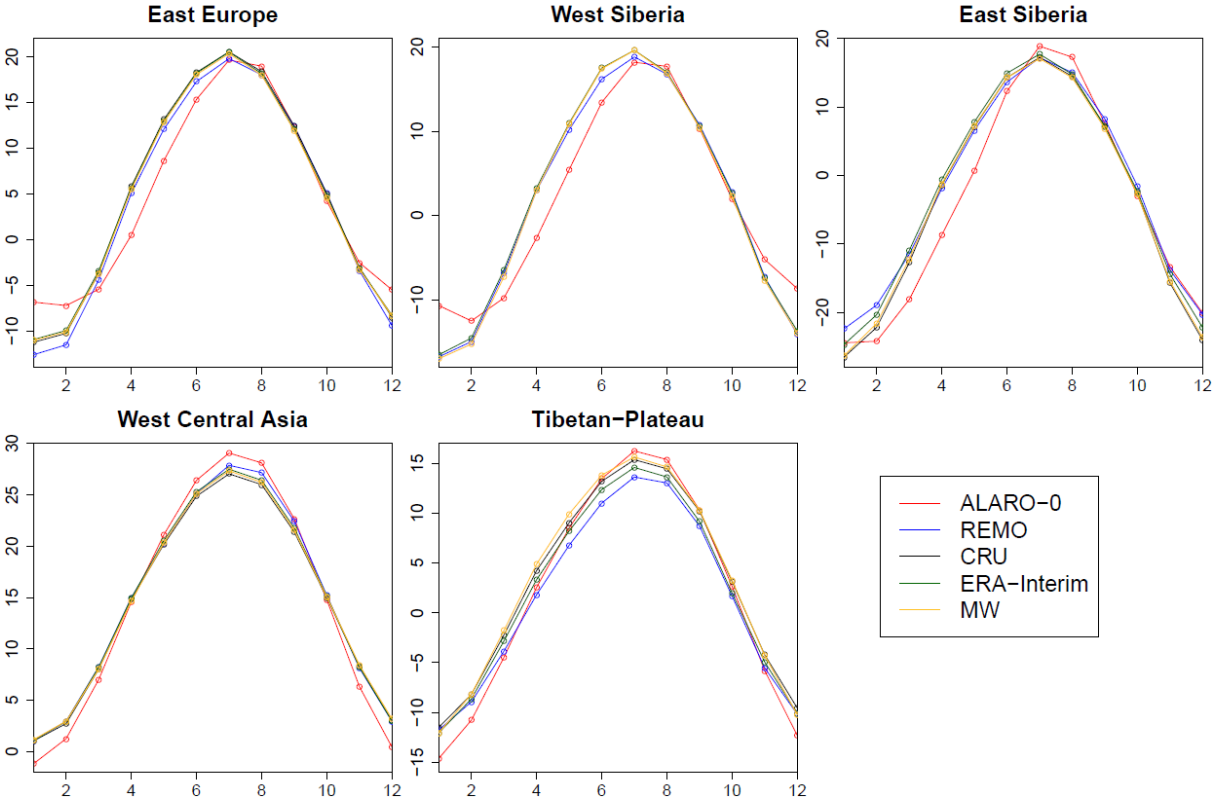

Fig. A1: Annual cycle of the mean temperature (°C) over different subdomains.

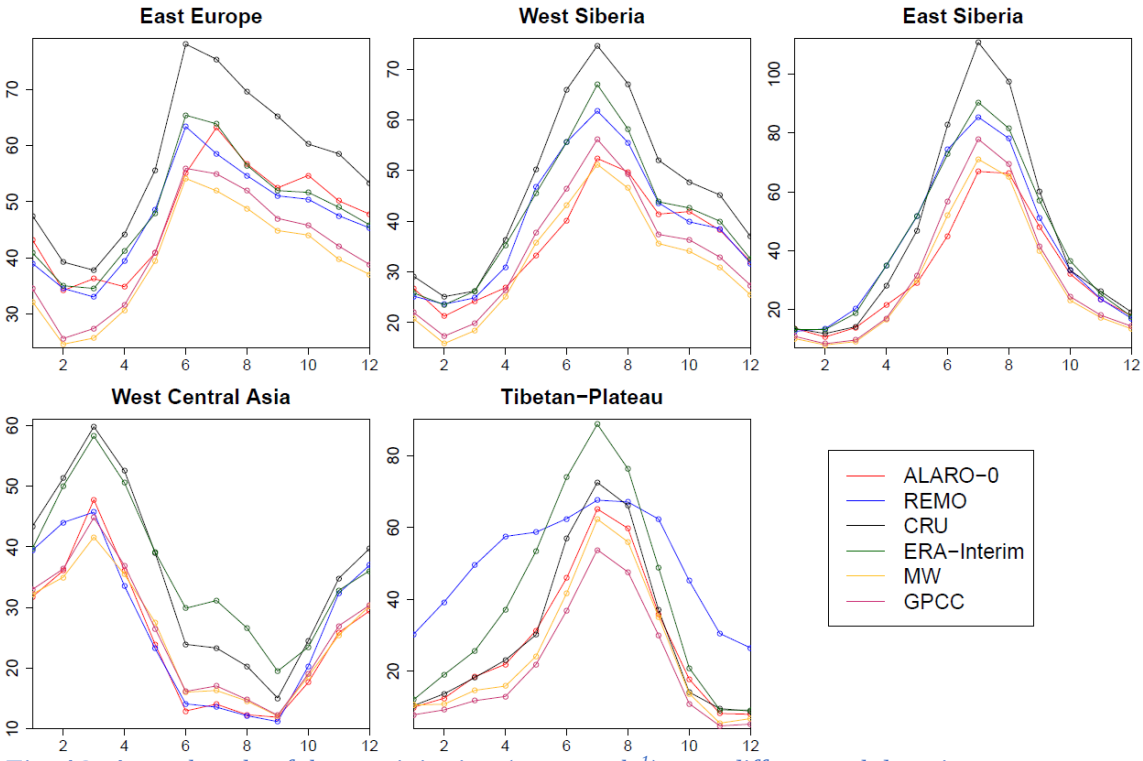

Fig. A2: Annual cycle of the precipitation (mm month^{-1}) over different subdomains.

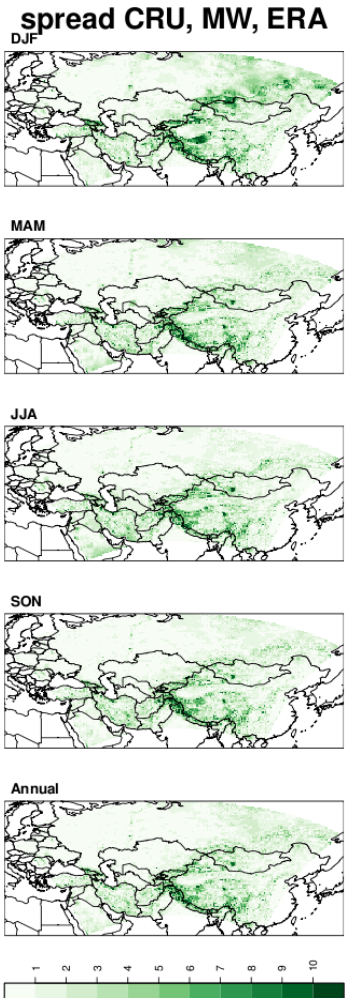

Fig. A3: Spread in mean temperature between the gridded datasets CRU, MW and ERA-Interim.

General Comments

In this paper the authors present the results of an evaluation conducted over the CORDEX Central Asia domain for two different RCMs: REMO and ALARO-0. Comparing climatological seasonal and annual means obtained from two simulations covering the period 1980-2017 against gridded observational data-sets, they aim to assess the reliability of the models for the region of study, setting the basis for their use for future climate projections. The paper complements the results of other studies on RCMs for the same region, and is believed to be interesting for the regional climate modeling community. Nevertheless, in its current form the paper suffers from a series of major issues that need to be carefully addressed before it may be considered for publication for Geoscientific Model Development. In general, the quality of the paper is not very satisfactory. The text and the structure of the manuscript need a thorough revision, since information is many times not very clearly expressed or confusing. The presented analyses are too generic and not at all exhaustive. Explanations for evinced models behavior are often hypothesized without any appropriate investigation. Further, I think that different sources of uncertainty such as the error related to the use of different observational data-sets are not properly considered. I discuss the mentioned issues, together with additional ones, in more details below:

Specific comments

- The presented analyses are neither exhaustive nor accurate enough for a proper evaluation study. In particular, the analyses of the spatial correlation and of the spatial mean calculated over the entire domain are not very useful. First of all, the evinced conclusions for the mean of the spatial biases calculated over the entire domain might simply be the results of some compensating effect and could vary significantly from one area to another. At the same time, given the heterogeneity of the domain of study, spatial means and correlations calculated over the entire domain can hide model limitations specific to single regions characterized by different physical phenomena. Determining and understanding possible model limitations is one of the final goals of models evaluation and serves as the basis for models development. For these reasons, a quantitative analysis of model performances per sub-regions is therefore required.

We agree that there might be some compensating effects due to the spatial means over the large domain. In order to improve our analysis we will add a section evaluating the RCMs over subdomains that are defined by the IPCC6 report (Iturbide et al., 2020) and that are situated in the CAS-CORDEX domain.

- In the text there is lot of confusion between the different sections and their contents, with discussion performed in the results section and some of the results commented in the discussion part. Also, the authors discuss several variables not in the appropriate subsections. One example is the subsection with the discussion on precipitation results, where the results of temperatures are also partly discussed.

To account for this comment we will rearrange the text in the results and discussion section to improve the readability of the text.

- The authors somehow considered the effect of different observations on the comparison of the maps of climatological biases, as well as for the spatial mean calculated over the entire domain. Nevertheless, also the analysis of the spatial correlation, the ratio of standard deviation and the RMSE should take into account the effect of different sources of uncertainties, among which one of the most important is certainly the effect of different observations. In this context, sub-regions analyses assume even more importance. Additionally, other uncertainties could play a big role for the different regions, such as for example the effect of different boundaries. What happens when these sources of uncertainties are considered? The authors should acknowledge the possible effect of different uncertainty sources and all their analyses must at least take into account the effect of the observational uncertainty on the considered metrics.

In order to visualize the uncertainty in data of gridded observational datasets we will add graphs to the manuscript with curves that show the differences between the gridded observational datasets for the annual cycle of the different subregions. The spread of the curves of the different gridded datasets can be considered as a measure of uncertainty.

It is indeed true that the positioning of the boundaries might have an impact on the climate experiments (Rummukainen, 2009), but it is not the aim of this paper to investigate the effect of the domain choice on the resulting RCM data. This work is undertaken within the CORDEX framework which provides guidelines on domains, resolution,... in order to enable RCM intercomparisons between different modelling groups. Therefore, we used the CAS-CORDEX domain as described by the CORDEX project for our model experiments. Although running the same RCMs over different domains would be interesting, it does not fit the aim of our study that frames into the AFTER project. Moreover, it is impossible to realize such an investigation on a short timescale as it would necessitate writing new proposals to obtain computing time on the Tier-1 HPC infrastructure.

- The authors conducted their evaluation considering a single observational data-set for each variable. Then, they basically discussed in each case whether and when the model bias was related to the poor quality of the reference observations, by comparing these with two or three additional data-sets. I am quite critical with the approach they used. In fact, a simple comparison of three or four gridded observational data-sets does not allow to determine the best data for the different regions of the considered domain. For doing this a more robust analysis is needed, considering the initial observational stations of each data-set, their number, their precision and the uncertainty related to the employed interpolation methods. On the other hand, what the authors can do, given the considered data-sets, is to evaluate and take into account the reliability of the given observational data-sets for each point of the domain, by calculating for example the spread of the different observations. Instead of determining whether evinced model biases are due to the reference observational data-set (that in my opinion is not possible to conclude for all the points of the domain, given the available data), the authors should compare models results with the available data-sets, and then discuss whether those biases are within or outside of the range of the observations. In this way they could be able to affirm whether any conclusion on model performances can be drawn for a considered area.

We agree with the reviewer that the spread on the observational datasets is relevant when evaluating the performance of the RCMs. Therefore, we will add maps with the spread between the gridded datasets instead of using Fig. 10 and 11. Additionally, we will add graphs to the manuscript showing the annual cycle of each gridded dataset and each RCM for different subregions. The difference between the curves of the different gridded datasets shows the spread between the different gridded datasets which can be considered as a measure of the observational uncertainty and provides evidence of the performance of the RCMs.

- The authors only investigate climatological values, focusing on the mean bias and on spatial variability. I think that they should be more specific on the choices they made, discussing at least why they focused only on seasonal and annual means and why they did not decide to tackle temporal variability and the seasonal cycle. In particular, I would suggest to add some analysis on the mean seasonal cycle, since the authors claim in their manuscript that some of evinced model biases might likely be related to a wrong simulation of it.

We indeed focused too much on the spatial variability. We agree that the temporal aspects can not be fully understood with the current figures in the manuscript, therefore we opted to add graphs with annual cycles based on monthly data for different subregions.

- The authors focus their analyses on four variables, but then they discuss the results only for temperature and precipitation. Why the discussion is not conducted consistently among the different variables? Additionally, why are the analyses of tmin and tmax not carefully conducted

in the same way as for precipitation and temperature, considering different observational data-sets?

We needed the minimum and maximum temperature together to bring the story about the limited diurnal cycle of the RCMs. That is the reason why we decided to split the discussion only in temperature and precipitation which indeed is different to the results section that had four subsections. We will merge the sections of minimum and maximum temperature in the results in a general section about the diurnal temperature range so it is clear that the different variables should be interpreted together to understand the processes later on in the discussion. In the discussion we will use the same structure of subtitles.

The evaluation of Tmin and Tmax is not conducted in the same way since the observational data was not available for all gridded datasets. The Matsuura and Willmott dataset of UDEL does not contain data about the Tmin and Tmax or the diurnal temperature range.

- The effect of the boundaries on the different variables can only be estimated by performing different simulations changing the boundary conditions. The authors should take into account this point whenever they claim that errors in the boundaries are the cause of evinced biases. They can eventually be able to (only partially) support these claims only by performing additional simulations with different boundary conditions.

As mentioned before, it is out of the scope of our research and the manuscript to do an in depth study of the effect of the boundaries due to the aim of the use of the CORDEX domain, the restricted computing time and the goals of the AFTER project which were the driver of these CAS-CORDEX simulations.

- In many cases the presented analyses are very superficial and most of the raised conclusions are mainly hypothesized without a proper demonstration. Additionally, sometimes the authors simply use the maps of the bias to interpret the results of the spatial means. Why should that be interesting and how such an analysis might help in evaluating models results? More in depth analyses are required, including the already mentioned investigation of the seasonal cycle and a quantitative comparison of model results and observations per sub-regions. Every hypothesis on the possible reason of evinced biases should be effectively supported by specific analyses or by bibliographic references.

As mentioned above, we will add a section with subregions that are lying within the CAS-CORDEX domain. We agree that an annual cycle based on monthly data improves the evaluation and insights. We will check which statements are not substantiated enough and we will add evidence that is forthcoming out of the added figures or we will refer to other scientific articles where needed.

- In the manuscript there is a tendency of justifying the bias of the model with the poor quality of CRU. For precipitation, for example, the authors state that CRU underestimates precipitation values: in this case, why do not you perform the comparison against the GPCC as reference then?

GPCC does not contain temperature data. Since it was important for us to refer for each variable to the same reference dataset in order to compare the performance of the different variables, we took CRU as a reference. By adding the analysis of the annual cycle over the subregions it will be easier to compare the RCM outcomes with the different datasets directly.

- I would be more careful stating that evinced results are in the range of the ones obtained for other studies and that indeed the models can be employed for climate projections. You affirm this only considering the reference of Kotlarski for Europe. Additional studies for more regions should be considered. Still, the authors must acknowledge the fact that extremely large biases are present over extensive parts of the domain. For these, either any conclusion on model reliability can be drawn due to high observational uncertainties or model results can not be considered very trustworthy.

We agree, for some parameters significant biases are present over parts of the domain for some seasons. The ALARO-0 RCM has a large positive temperature bias in winter over the northern part of the domain. The REMO model has difficulties in reproducing the observed precipitation patterns over the orography of Central-Asia. We agree that the biases observed in this study should be kept in mind when presenting future projections. We find it therefore important to publish an exhaustive evaluation study. In this evaluation study we saw that the main patterns are modelled correctly and therefore we concluded that we can move on towards climate projections. We will add to our conclusion that these large biases should be kept in mind when looking to the future projections. Additionally, to deal with the biases in impact studies, several bias adjustment methods have been tested within the AFTER project and the most suitable method will be applied before simulations for impact studies are done with these climate data. It is not in the scope of this evaluation study to explain the details about bias adjustments and impact modelling but to avoid misunderstandings we will add that bias adjustment is one of the possibilities when mentioning that the RCMs can be used for future projections.

In other scientific publications where models over the CAS-CORDEX domain were run there are as well large biases over certain parts of the domain (Ozturk et al., 2012; Ozturk et al., 2016; Russo et al., 2019) and even for RCMs run over subregions large biases were found (Wang et al., 2020; Zhu et al., 2020). There are not a lot of scientific articles to compare our results with and to refer to, however in the meantime some new studies are published with model evaluations over a subdomain of our domain and we will refer to them in the updated manuscript. We will thus rewrite the discussion and refer to more scientific articles.

Minor Comments

- lines 35-37: How large are these ensembles? what about ensembles for the other CORDEX regions, such as for example North America?

These large ensembles consist all out of more than ten GCM-RCM combinations. For example, the ensemble of the EURO-CORDEX domain consists of 14 GCM-RCM combinations; 18 GCM-RCM combinations are available for CORDEX-Africa. North America contains as well a large ensemble of 13 GCM-RCM combinations for the 0.22° resolution but we did not want to list all the different CORDEX regions and the number of GCM-RCM combinations. In our submitted manuscript we mentioned EURO-CORDEX, CORDEX-Africa and MED-CORDEX but NA-CORDEX has indeed more GCM-RCM combinations at the 0.22° resolution. In the revised version we will therefore replace MED-CORDEX by NA-CORDEX. A detailed overview of the available ensembles over the different CORDEX regions can be found at the official CORDEX website: https://cordex.org/ and for each CORDEX domain there is a tab on this website with more information or a link to the website of that particular CORDEX domain.

- line 51: the term "validating" is normally considered not very appropriate when comparing climate models and observations, in particular in cases like this one, where large uncertainties in observations are present. "evaluating" would be a more appropriate term.

As suggested, we have changed "validating" to "evaluating".

- line 56: Same as above. Replace validation with evaluation everywhere in the text.

As suggested, we replaced "validation" with "evaluation" in the text.

- lines 62-63: delete "...that are sparsely populated" since it is repetitive (you already said that in the first line of the period).

As suggested, we removed it.

- line 65: "more extreme values": more extreme than what? just use "extreme values"

We agree and we changed it in the text.

- lines 67-68: The comparison against different observational datasets is useful only to address the reliability of observational datasets and does not help solving the problem of the lack of an ensemble. Please reformulate.

It is reformulated.

- lines 70-71: Similarly as expressed in my major concerns, you cannot directly prove that similar biases in the two models are due to observational errors.
In principle, uncertainty in the observational data-sets allow to say that over certain areas the observations are more or less reliable and whether robust conclusions can be drawn in this case. Please reformulate this part.

It is reformulated.

- line 80: complemented by

It has been corrected.

- lines 83-88: I think that this part would be more appropriate for the introduction rather than for the methods.

We agree, the text has been changed.

- line 106-107: "The outer domain consists of the inner domain plus a coupling zone of eight grid points in each direction.": This holds true for both domains, right? eventually specify.

Indeed, this is true for the domains of both RCMs. We specified this in the text so it is clear that we refer to both RCMs with this sentence.

- Fig.1: Where did the authors take the information on the topography from? the upper limit of the colorbar of 3000m seems not reasonable for the area.

The figure shows the values of the topography used in the regional climate model REMO [GTOPO30 global digital elevation model (DEM) 3 https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30?qt-science_center_objects=0#qt-science_center_objects]. The explanation is added to the figure's caption.

We have increased the upper limit of the colorbar to the upper limit of the orography within the study area.

- line 131: what is the vertical extension in meters of the domain of study for each of the two models?

For REMO, with 27 levels, the top is approximately at 25 km height. The top of the uppermost gridbox is set equal to 0 hPa, but in reality the midpoint of the uppermost gridbox is ~25 km. ALARO-0 uses a vertically staggered grid and the top of the uppermost gridbox is also set equal to 0 hPa. The midpoint of the uppermost gridbox is situated at 67 km for a standard atmosphere.

- lines 139-140: Correct into: "...at the boundaries, up to the 31st of December 2017."

We corrected this sentence.

- lines 143-147: Is there any reason why in the case of ALARO-0 one year can be considered enough for spin-up with respect to the 31 years considered for REMO? Please specify.

Both RCMs are using a different soil model. The soil model used for REMO is using five layers with a mean rooting depth up to 5.7 m (Kotlarski, 2007), while there are only two layers in the ISBA model for ALARO-0. One year spin-up is enough for ALARO-0 since different variables reach their equilibrium after maximum one year. Most soil properties find their equilibrium after about one month. To reach an equilibrium state for the soil temperature and soil moisture, a warm spin-up period of ten years instead of thirty years was used for REMO. We will correct this in the text.

- lines 149-150: This sentence, in the way it is expressed, is not properly correct. In fact, you do a comparison of model results only against the CRU, while then you compare the different observations among them. Please better reformulate this sentence according to the comparison you will decide to perform.

We decided to add annual cycles of the different datasets and RCMs, thus this sentence should not be changed since in those new graphs the results of the RCMs are compared with the different datasets.

- lines 151-152: Again, given your analyses you can not tell whether the bias of the models is due to the observational uncertainty. What you can eventually say is that high uncertainties do not allow to draw robust conclusions.

It is reformulated.

- lines 160-165: I do not manage to find the reference of New et al. 2002 in your paper. Are not there any more up-to-date publications discussing problems of the latest CRU releases? Also, do not you think that the New et al. 1999 publication is more general and it might also apply to other observational datasets rather than simply the CRU? Please consider that all your considered data-sets are somehow characterized by uncertainties (Flaounas et al., 2012;Gómez-Navarro et al., 2012).

We checked and updated our references in this part of the text. Recently a new paper for the CRU data was published (Harris et al., 2020) and we updated our text taking this paper into account. Indeed, New et al. (1999) is rather describing general features about gridded datasets but they do focus on the first versions of CRU, that is why we mentioned this reference as well in the section about CRU. We agree that it is better to refer to more recent and concrete papers for the CRU dataset. Additionally, we will add a sentence in the general part about the reference datasets taking into account Gómez-Navarro et al. (2012). The study of Flaounes et al. (2012) (about the ECA&D gridded dataset over MED-CORDEX) is not general enough to be relevant for our text.

- lines 168-171: what about quality of UDEL for other variables than precipitation?

We added information about the variable temperature.

- lines 176-179: Please, make clear that Hu et al. 2018 only investigated the most central part of your domain of study. Also, the same study states that GPCC underestimates all seasonal means, not only but especially in spring.

Adaptations have been made in the text as suggested.

- lines 181-186: The original resolution of ERAInterim is not 25 Km but approximately 80 km. If you used the data provided by the ECMWF at 25 Km, be aware that these are interpolated data. Please specify this in the text.

As suggested, the explanation has been added to the text and adapted in Table 1.

- lines 181-186: an additional question concerns your choice of using ERAInterim data interpolated at 25 km: why you do not directly download ERAInterim data already onto a 50 Km grid?

We had ERA-Interim available at 25 km on our HPC infrastructure and a projection to the 50 km grid results to the same. The new graphs with annual cycles are not produced at the 0.50° resolution but at the resolution of each dataset and 0.22° for ALARO-0 and REMO.

- lines 186-188: First of all avoid saying initial errors in the boundary conditions, since it generates confusion. Then, how should the comparison of temperature derived from ERAInterim with the one of the models help you determining what is the effect of errors in the boundaries? The only way to assess the effect of the boundaries on the RCM results is to drive the same simulations with different boundaries.

We agree that it is confusing. To be certain about the effect of the errors at the boundaries, other boundaries should indeed be applied. We have deleted this part in the text.

- lines 189-190: The outputs of an RCM are dependent (but not univocally determined) on the values of several variables with which the model is forced at its boundaries. These variables will have an effect on several model variables. The temperature of the model is not only dependent on the values of temperatures provided at the boundaries, but other variables play a role. The same holds true for precipitation. If the model is forced with wrong temperatures it is very likely, at least from a theoretical perspective, that both model temperature and precipitation will be both badly reproduced.

We agree, the text at this line was deleted.

- lines 198-199: First of all UDEL and CRU have the same 0.5 degree resolution. Also GPCC is available at such resolution. Reformulate this period in a more accurate way, considering the fact that the "upscale" is only necessary for the models outputs.

We have changed "upscale" in the text as was suggested. The annual cycle graphs were created using the highest resolution of each dataset (0.50° for CRU and UDEL, 0.25° for GPCC and 0.80° interpolated to 0.25° for ERA-Interim).

- lines 207-209: reformulate this period.

We reformulated this part in the text.

- line 219: seasonal means of

We corrected this.

-line 227: what do you mean by limited bias? better specify.

We reformulated this part in the text.

- lines 228-229: First of all, you start discussing annual means but you put the relative figures at the bottom row of your image: move them up. Then, in my opinion, according to the scale you use in your plots, it seems that in both cases the absolute bias exceeds 3C over a very extensive part of the domain and not only over mountainous regions. Maybe the scale you are using does not help to clearly distinguish which areas are above or below a certain threshold. Try to change your scale.

We agree that the maps of annual means have to be placed at the top of the figure. It is indeed difficult to see the difference between each degree on the figure. We will change the scale.

- lines 229-230: Also the REMO exceeds the 3C range, in particular in winter. Please reconsider your sentence.

We included this REMO temperature bias a bit further in the text where we discuss the biases in the mountainous regions and say that REMO has a warm bias in winter over the Altai region. We agree that this might have been confusing and as suggested, the warm bias of REMO that exceeds 3 °C in winter over the north-western part of Mongolia has been added at this particular location in the text.

- lines 230-231: not totally correct. In fact, the biases ,when considering the entire domain, are particularly pronounced for ALARO-0 mainly in spring, over the northern part of the domain. In winter the most pronounced bias seems to be the one of REMO for north-western Mongolia. For summer and autumn the biases for the two models present a very similar range. The same holds true when considering only the eastern half of the domain. Reformulate this part.

We reformulated this part as suggested.

- lines 231-233: Actually you should really emphasize that the two models seem to have a completely different pattern of the bias of temperature in winter: one shows a bipolar behavior between North and South, while the other between East and West, with a peak in warmer simulated temperatures over north-western Mongolia. I think that it would be really important for the authors (and a very nice opportunity) to better investigate the causes of the two different behaviours. This could give us some clue on model limitations in the simulations of temperatures over the region, that seems to be a general issue for climate models.

As suggested, we will emphasize this different behavior of the models in the text. By including an additional subsection showing the yearly cycle of both temperature and precipitation of the observational datasets and model output over subdomains the reader gets more insight into these bias patterns.

- lines 233-234: What do you want to evince from this? why Scandinavia and not another region? Also, how is the bias similar in the two cases?

We moved this information to the discussion section. The climate in Scandinavia is similar to the climate in the northern part of the CAS-CORDEX domain. The reason why in both regions a warm bias is obtained for ALARO-0, is probably linked with a process that occurs in regions with a subarctic climate and not somewhere else. Deviations in snow related processes might explain the warm winter and cold spring temperature biases in the northern part of the domain and therefore we will add some additional information in the discussion part about this feature. We are currently investigating this.

- lines 238-239: Important biases are present in MAM also for REMO, for some regions such as the Western fringes of the Tibetan Plateau. Also, for both models biases exceed 3C over a large part of the domain in MAM. Reformulate.

As suggested, these sentences have been reformulated.

- lines 239-241: What do you mean by limited? you mean that biases are not very pronounced in summer? reformulate.

We reformulated this sentence as suggested.

- lines 239-241: also for REMO there are warm biases, even though they are inherent to a smaller portion of the domain, in particular with respect to ALARO. Be more precise.

We reformulated this sentence as suggested.

- Fig 2: Beside my previous comment on the figure colorbar, the quality of the image could be further improved by reducing white spaces in between rows and moving the names of the seasons on the left side of the figures. Additionally, units should be added to the colorbars, that should also be moved: the colorbar of the bias should be positioned in between the two columns for the bias of REMO and ALARO.

We decided not to change the location of the names of the seasons in the figure. By placing the names to the left side of the maps the maps would become smaller in order to fit the page. We want to present our figures as large as possible and that is why we structured it in this way. We will add the units to the colorbars and place the colorbar of the bias at the right side of the figure.

- lines 248-249: The mentioned gradient is not very clear, in particular in summer.

We removed this statement.

- line 249: "The outcomes of both RCMs for the mean temperature agree well with the CRU data in autumn (SON)": That is not totally true. In fact, performances of REMO in terms of simulated seasonal climatologies are very similar for autumn, but also for spring and summer.

We reformulated this sentence.

- lines 254-255: what do you mean by "should be placed in perspective"? in which perspective? please reformulate this period.

We reformulated this sentence to make clear that the uncertainty in observational gridded datasets is known to be larger at locations in mountainous areas.

- lines 258-259: "it is clear from Table 2 that the strong cold bias during spring in the north for the ALARO-0 model has a larger negative impact on the spatially averaged bias than the warm bias during winter": I would avoid talking about "negative impact" of the bias over some region on the calculation of the spatial mean bias. Instead, you could say that the spatial bias is largely influenced by the pronounced negative/positive bias over specific regions.

Thank you for the suggestion, we reformulated this sentence.

- lines 264-267: "However, the biases during summer are ... due to the smaller spatial variability in temperature during summer". I think that this period is not very clear and needs to be reformulated, eventually considering additional analyses supporting your conclusions. First of all in summer, in the observations, you have less spatial variability (more accurate than smaller spatial range) than in the other seasons. This is evident from the figure, even though it would be nice if you could support such conclusion with a more quantitative analysis of the CRU spatial variability. Additionally (and most importantly), in your analyses you do not effectively demonstrate that a lower correlation is due to a lower spatial variability in summer. Why can it not be simply due to a worse agreement in the spatial variations between the models and the observational dataset?

We agree that the sentence at line 264 is confusing and does not add any value, therefore we decided to remove this sentence. We have changed "smaller spatial range" into "less spatial variability". We will not include a more quantitative analysis of the CRU spatial variability to keep the document as concise as possible and since it is already visually clear from Fig. 2 that the spatial variability is smaller in summer. From Fig. 2 it is visually clear that the biases are lower in summer compared to winter and autumn, thus we assume that the lower spatial variability in summer is the reason for the lower correlation and not the worse agreement between the models and observational dataset.

- lines 276-277: that is exactly one of the reasons why it would be better to consider the analyses per sub-regions.

We agree and will take into account a subregional analysis.

- lines 290-291: I think that your explanation on the reasons of a more negative bias for TMIN than for T2 is not exhaustive. Additionally, this needs to be moved to the discussion part.

We will move this to the discussion section, where we can explain it exhaustively.

- lines 297-298: "Following the main trend..": confusing, reformulate.

We reformulated this sentence.

- lines 299-301: "The warm minimum temperatures of the RCMs indicate that they underestimate the coldest diurnal temperatures or that the observational CRU dataset overestimates them." There are several issues in this period. First of all you need to reformulate your sentence because it is not the minimum temperature of the model that underestimates observation values but rather the model itself. Also, if the minimum temperatures are warmer than observations, it means that the model overestimates (and not underestimates) the coldest diurnal temperatures. Finally, from the comparison of model results against CRU you can only affirm that the models underestimate minimum diurnal temperatures. You can not prove that the observations overestimate them. The fact that CRU might overestimate them is a possibility, but still is not inherent to the behaviour of the model (nor it is evident from the figure you are commenting).

We reformulated this sentence according to the suggestions.

- lines 312-313: "except for the summer": why except if your are talking about annual values?

We agree and reformulated this part of the text.

- line 315: less good than what?

We reformulated this part of the text and moved it to the discussion section.

- line 323: you do not need to specify that temperature is a variable here

We agree and, according to the remarks that were made for minimum temperature, we moved this sentence to the discussion section.

- lines 323-324: You need to reformulate this sentence. In this case you have to specify that the negative TMAX bias is particularly remarkable in spring for the northern part of the domain, and also, to a less degree, in summer. In winter some other less extended parts of the domain, such as the north-eastern part, show a colder bias than REMO. In Autumn results are more similar between the two models.

We agree and we reformulated this text part.

- end of line 326: the cold bias

We corrected the typo.

- lines 326-328: Fig. 4 shows minimum temperatures. Then, how can we deduce from this figure that the bias in TMIN is due to maximum temperatures? please better explain and eventually reformulate this period.

We referred to the wrong figure, it should be Fig. 6. This sentence is describing what was earlier mentioned: "specify that the negative TMAX bias is particularly remarkable in spring for the northern part of the domain". We moved the sentence up and rewrote it a bit so it is clear what we are trying to say.

- line 342: "This means that ALARO-0 fails to reproduce the low nocturnal temperatures": This belongs to the discussion on minimum temperatures. Additionally, the model still fails in simulating warmer temperatures, despite the smaller bias when compared to TMIN.

As suggested, we moved the last paragraph of this section to the discussion section and we explain more clearly that ALARO-0 fails to reproduce temperature in general (including mean, minimum and maximum temperature) in the northern part of the domain.

-lines 344-346: You should discuss about minimum temperature in the appropriate section.

As suggested, we moved the last paragraph of this section to the discussion section.

- When you comment the maps of the bias, try to discuss the different seasons from up to down, consistently with the figures.

We agree.

- lines 364-366: This part should be moved to the discussion section.

As suggested, we moved this text to the discussion section.

- lines 365-366: why however? also, you did not discuss until this point the uncertainty of CRU: how can you claim that the reason for the wet bias is due to the observations?.

We agree that "However" at the beginning of this sentence is not suited here. It was not our intention that this sentence was interpreted as a shortcoming of the CRU dataset since this is the results section. We wanted to express that it is known from the observations that the amount of precipitation is low in certain regions as seen in Fig. 8 (< 5 mm/month), not that CRU contains precipitation amounts that are too low (this follows in the discussion). We reformulated this sentence, to overcome the confusion.

- lines 370-372: By whom is the bias turned into something else in summer? and how?

We reformulated this sentence to make it clear that we talk about summer, when the East Asian Monsoon takes place.

- lines 372-376: It would be nice if you could perform the analyses of the seasonal cycle to support your conclusions. This would make your evaluation more complete and exhaustive, while at the same time allowing to effectively confirm or deny your conclusions.

Thanks for this suggestion. We agree and did an analysis of the annual cycle over multiple subdomains.

- lines 390-391: "The dry biases for ALARO-0 in Table 5 are thus caused by the simulation of systematically less precipitation than the precipitation amounts in the CRU data.": Reformulate. It is obvious that if the model underestimates precipitation, it simulates less precipitation than observations.

We reformulated this sentence. We intended to say that there is no region that has a strong dry bias which is compensated with a wet bias in another subregion. This differs from the finding of temperature where the strong warm bias in the north is partly compensated by a cold bias in the southern part of the domain.

- line 393: systematically

We corrected the typo.

- lines 392-394: "The lower accuracy of simulated precipitation is due to the fact that precipitation is less systematic affected by land cover and topography compared to temperature": First of all that is quite a strong assumption given the extent and heterogeneity of the domain you are considering. Additionally, you did not perform (at least it is not reported in the paper) any analysis that supports your conclusion.

We agree, we did not perform an analysis on this topic but it is known that it is harder to simulate the spatial pattern of precipitation compared to temperature (Kotlarski et al., 2014) due to the reason we mentioned.

- lines 400-404: This is incorrect. In fact Russo et al. 2019 showed that uncertainty in observations is high over the north-eastern part of the domain, not that CRU overestimates the diurnal temperature range over the region.

We agree and will reformulate this text part.

- line 404: Why hence?

We agree that this is an incorrect cause-consequence structure and we will reformulate it.

- lines 404-406: Again, how can you surely state that the model underestimates values of the diurnal temperature range due to higher observation values?

We will rewrite this part.

- lines 407-408: why Czech Republic? what happens in other regions?

We agree that it would be better to refer to literature over Central Asia instead of referring to literature over EURO-CORDEX where ALARO-0 and REMO were already evaluated. We will refer to Russo et al. (2019) who obtained similar findings.

- lines 420-422: This is just an assumption that needs to be proven. Models develop their answer that is, to a certain degree, independent from the boundaries. To test your hypothesis, one easy experiment that could be conducted is to use different boundaries and compare the results.

We agree and we will remove this.

- line 425: "They related this warm bias already to shortcomings in the simulation of snow": this means that they explained the bias differently than with the boundary effect as you explained in the lines from 423 to 425.

Ozturk et al. (2012) explained the bias indeed with a shortcoming in the simulation of snow cover. We will remove the part about the boundary effect.

- lines 430-431: "Hence, we conclude that the warm forcing is the main reason for the warm bias over Eastern Russia during winter.": I further have to highlight that you can not make such conclusion, until you do not test different boundaries.

We agree and we will remove the part about the boundary effect.

- lines 435-436: As before, it would be nice if you could do the analyses of the seasonal cycle since you mention it for the interpretation of your results.

We agree and we will add annual cycle graphs as mentioned before.

- lines 440-442: How the fact that for Belgium there is some correlation between warm bias and cloud cover representation could explain the same for northeastern CAS. You could do some analyses on cloud cover to support your conclusion.

We mentioned this study over Belgium since it is the only study that investigated the relationship between temperature and cloud cover for ALARO. We agree that we cannot draw strong conclusions from this and that this previous paper only gives a clue that cloud cover might be one of the reasons why the temperature is not well estimated. Cloud cover is thus only one out of the many possible reasons, which should be further investigated. In the meantime we did some analysis on cloud cover and we will include our findings to the new version of the manuscript.

- lines 443-445: "Both could be due to too much cloud cover": according to whom? In theory it could be due to any reason.

We agree and we investigated this further to say something about it in the discussion.

- lines 448-451: These considerations are important: it would be nice to put them in a more objective context. Additionally, you say that New et al. show that CRU underestimates temperatures for Russia. Then you talk about Western Russia. If you state that temperatures from CRU are not good for Russia, then they can not be good for a part of it and bad for the rest. Reformulate.

We reformultated these sentences based on the additional analysis over the subregions.

- Fig. 10,11: To make the discussion easier I would suggest to plot the maps of the differences between different observational data-sets together, using the spread of the observations among the different data-sets. In this way you can easily know which areas are more reliable and which are not.

We agree and produced new figures.

- lines 465-467: the less reliable observations do not explain the bias, rather they do not allow to draw any conclusion.

We reformultated these sentences.

- line 471: "...winter and overestimate it during." During what?

The word "summer" is missing, we added it to the text.

- lines 482-484: what happens when you compare ALARO-0 with the other data-sets over the entire domain?

The precipitation of ALARO-0 is for most grid points within the range of the different gridded datasets during the different seasons. When averaging over the complete domain, then the output of both RCMs is within the range of the spread between the reference datasets for the different seasons. However,

there are some subregions where the precipitation of ALARO-0 and/or REMO is lower or higher than the observational spread for a specific season. For example both RCMs slightly underestimate precipitation in summer over West Central Asia. We will add this information in the updated manuscript.

- lines 484-486: you mention two gridded data-sets: to which data-sets are you referring here? please better specify.

We are referring to GPCC and MW, these are observational gridded datasets. ERA-Interim is a reanalysis product, so we do not refer to it as an observational gridded dataset. We reformultated this sentence and we will make sure that this is clear throughout the complete manuscript.

- lines 486-489: again, you can claim that the bias is relative to the employed boundaries only performing a new simulation with different boundaries. Also, how can you be sure that ERA-Interim overestimates specific humidity?

We will reformulate these sentences since it was not intended to say that the boundary conditions of ERA-Interim affected our results. We just wanted to point to the similarities between the ERA-Interim data and the output of the RCMs. We agree to remove the suggestion of the overestimation of the specific humidity as we did not investigate it.

- lines 489-490: You are claiming this from the field means I guess. I think that plotting the maps of the bias of the models against all the different observational data-sets might help the discussion of your results.

We claim this based on Fig. 8, 11, S1 and S2 where the spatial patterns between ERA-Interim and REMO are visually very similar, while the patterns of ALARO-0 are similar to GPCC and MW. We agree that it can help to plot the maps of the bias of the models against the different observational datasets, however this will make the manuscript long.

- lines 489-490: How do ERA and REMO parameterize precipitation since you mention that they do it differently? Specify.

For REMO these specifications are included in Table S1. We will refer to this table at the end of this sentence and we will add that ERA-Interim uses a convection scheme modified from Tiedtke (1989) by Bechtold (2008; 2014) and the cloud scheme is based on Tiedtke (1993) with modifications made made by Forbes and Tompkins (2011), Forbes et al. (2011) and Tompkins et al. (2007) (https://www.ecmwf.int/en/research/modelling-and-prediction/atmospheric-physics). The similarities between ERA-Interim and REMO for precipitation are thus probably due to the fact that both use a modified scheme that is based on Tiedtke (1989). We did not further investigate this.

- lines 492-493: "This difference between ALARO-0 and REMO is related to the 3MT cloud microphysics scheme of ALARO-0": where did you demonstrate this?

We did not demonstrate or investigate this but it is an assumption since this is known to cause differences (Giot et al., 2016). We reformulated this statement so it is clear that it is an assumption that should be further investigated in the future.

- lines 496-497: Again, this is hard to affirm simply using three observational data-sets. The authors have to acknowledge the low number of observational data-sets. As I mentioned in many previous comments I personally think that it would be better to approach the differences between the observational datasets in terms of reliability rather than determining who is more correct.

We agree, we will mention the low number of observational datasets. We will reformulate the text so we focus on the reliability of the observations and the complications for our evaluation. It was not our intention that it looks like it is a research on which reference dataset is the best one.

- Fig. 11: In the colorbar of the bias, are units percentage? with respect to what? Please specify.

Indeed, the unit was wrong and we corrected it to the unit percentage. In Fig. 11 the precipitation of CRU is compared with the other datasets ERA-Interim, MW and GPCC. The relative values were obtained by dividing the difference by the value of CRU as already mentioned in section 2.4 Analysis methods. In order to make this clear in Fig. 11, we will specify this in the figure captation.

- lines 501-502: Not completely true. Specify that, as evinced from your maps, the wet bias in ERA (with respect to the other 3 data-sets) is only relative to the eastern part of the domain.

We reformulated this sentence as suggested.

- line 520: "that that". Correct.

We corrected this.

- lines 536-537: "REMO simulates the precipitation fairly well and ALARO-0 performs very well." How can you state that their performances are good?

The simulated precipitation of the RCMs is for most regions most of the time within the observational spread. We have clarified this in the manuscript.

- lines 539-540: "The warm temperatures obtained with REMO ... can be linked with the dry and wet bias in winter and spring respectively." Why and how can they be linked?

We agree that they cannot be linked without doing an in depth study on how they are exactly linked. We reformulated this sentence.

- lines 540-541: In which way the link between temperatures and precipitation should strengthen your hypothesis of a delay by REMO in the simulation of snow cover? Can you be more specific?

This was an assumption, we reformulated this sentence.

- lines 545-546: "The persistent warm bias over Pakistan and Northern India of both RCMs can be explained by the persistent underestimation in simulated precipitation over this region by both RCMs.": how can you state that given your analyses?

We agree that the warm temperature bias cannot be explained by an underestimation in precipitation. We reformulated this sentence.

- lines 547-550: You refer to the fact that your results are within the ranges of models for other domains, but then you only mention the results of Kotlarski et al. for Europe. You need more references.

We agree, we will add some papers that were recently published over parts of the CAS-CORDEX region.

- lines 562-563: That is arguable, given your analyses. How do you define an acceptable range?

An acceptable range is within the range of the observational spread. We will reformulate this so it is clearer.

- lines 565-567: You cannot state this, until you force the model with different boundaries and you conduct an analysis of snow cover (what you can eventually do for snow cover is to reference to the evidences from other studies).

We agree and we removed this sentence.

- Table 1: This table is not easily readable. Could you find a way to make the distinction between the different data-sets a bit clearer?

We will add a light gray background to the odd rows, so that the distinction between the information of the different rows is more clear.