Reply to the review of "Evaluation of regional climate models ALARO-0 and REMO2015 at 0.22 resolution over the CORDEX Central Asia domain" by Top et al. 2020

List of relevant changes

Dear Editor,

Thank you for giving us the opportunity to improve our paper. Based on the comments of the reviewer we have made the following major changes.

We describe the figures in detail for the description of the mean temperature in section 3.1. Since the subsequent figures in sections 3.2 and 3.2 have exactly the same format as the ones in section 3.1, we do not repeat the detailed description of the figures since the reader should be familiar with them by then.

We have made sure that sections 3.1, 3.2 and 3.3 all follow the same structure, i.e.

1. *description of the figure and the table (the table complements the figure) then discuss in this order:*
    1. *the performance of the annual mean,*
    2. *the performance of the seasonal mean,*
    3. *eventually the performance over steep orography*
2. *presentation and description of the Taylor diagram*
3. *presentation of the annual cycle.*

Additionally, we added the abbreviations of the subregions to the Taylor diagrams and we made the font larger of the values on the axis.

There is one citation Nikulin et al. (2012) that contained errors in the names, probably due to a wrong copy. We have corrected it.

In the supplementary material we corrected and added the identifiers of the used ALARO-0 data that is now available on ESGF.

Author response to the review of the Anonymous Referee

Dear reviewer,

Thank you for your helpful comments.

The authors improved the quality of the paper compared to previous versions. In particular, the presented analyses are now more appropriate for an evaluation of models performance. Nonetheless, the quality of the manuscript is still not sufficient to be granted publication. In particular, the description of the results is very poor, full of inconsistencies, not exhaustive and detailed and not consistent throughout the different sections. I invite the authors first to describe each of their figures, extensively providing all the details, and only then summarizing the text. Also, please check that you proceed in the description of the results in the same way for each of the figures. Below you can find a detailed list of some of the main issues of the paper. Be aware that errors are not limited to the mentioned examples, but all the manuscript needs a thorough review. The paper is at a good point and a thorough and patient review should allow for its publication.

• l. 57-58: by whom was the simulation performed?

The simulations were performed by the Met Office Hadley Centre (MOHC). The sentence is reformulated and a more suitable reference has been added.

• l. 61: A new paper by Russo et al. 2020 with COSMO-CLM is available for Central Asia, published on GMD.

We refer to it in the revised manuscript.

• l. 102: comparable with what? you probably mean among each others in a coordinated framework?

Indeed, this sentence can be improved. We now write: "In order to obtain simulations that allow for coordinated intercomparisons"

• Fig. 1: Specify in the caption that all points with orography higher than
3000 m are set with the same height of 3000m.

We specify this in the revised manuscript.

• l. 120-122: A regionalization was applied and not the subdomains. Reformulate.

We agree that it is confusing to use domain and subdomains, while we did only runs over one domain. It is indeed better to consistently use the word subregion instead of subdomain to refer to the smaller areas that we discuss in more detail. We now write: "In the present paper, the CAS-CORDEX domain was further subdivided into five subregions ...". Subdomain has been replaced by subregion throughout the text.

• l. 148-150: line repeated twice. Also, make example of which parameters you refer to.

Thank you for pointing it out. The repeated line has been removed. The parameters were already summed up in the sentence that follows: "These include sea surface temperatures (SSTs), surface roughness length, surface albedo, surface emissivity and vegetation parameters."

• l. 153, "to produce an equilibrium for the soil temperature and soil moisture": does not sound good. Reformulate into something like: to let the model reach an equilibrium state for …

We agree and now write: "... REMO was spun-up for 10 years to allow the model to reach an equilibrium state for the soil temperature and soil moisture ..."

• l. 156-157: it should be "were compared". Check for consistency of verbs tense throughout the text

Good point. The rule is here that the work we performed is written in the past tense, while the description of the data sets is in the present tense. We have changed one sentence: "... annually averaged values for temperature and precipitation were compared with different reference datasets..." We checked for consistency of verbs tense throughout the text.

• l. 161 "since all gridded datasets are characterized by uncertainties": I would avoid such statement, since it is quite obvious. On the other hand I would say that you consider different observational data-sets for the calculation of given metrics, for assessing the reliability of the results.

This is indeed an obvious statement. We now write: "A multitude of datasets were considered to assess the reliability of the gridded observational temperature and precipitation (Gómez-Navarro et al., 2012)."

• Section 2.3: make clear that you use CRU as reference and additional data-sets for assessing reliability of observations over different areas.

We made clear in section 2.4 that CRU is used as reference. However, we understand that it is worth to mention it in section 2.3 as well. Therefore, we added a sentence in 2.3.1: "In present paper, this data set is used as the reference while the spread of the data in all of the data sets is used to assess the reliability over the different areas."

• l. 166: was used

Thank you, we follow the rule mentioned under your comment on l. 156-157. We now write: "Monthly values of minimum, maximum and mean near surface air temperature and precipitation were used in the current study."

• l. 183-185: period a bit confusing, you need to reformulate. Basically you state that Hu et al. found that GPCC is better than CRU and MW for the inner part of Central Asia CORDEX domain. Than you say that precipitation is underestimated in montainous regions, but are you still referring to the inner part of the domain? Also, you state that GPCC underestimates precipitation in general: but for which region? Globaly? for Central Asia? do you have a reference?

We do indeed still mean that the overall underestimation in precipitation and underestimation in precipitation in mountainous regions by GPCC was found for the subregion defined by Hu et al. (2018). This period has been reformulated in the revised manuscript based on the suggestions.

• l. 199: as for GPCC, where ERAInterim generally overestimates precpitation? References?

Here, we meant that ERA-interim overestimates precipitation globally, especially over mountainous regions based on the reference of Sun et al. (2018).

• l. 216: by computing the difference between maximum and minimum

Indeed, this can be stated more clearly. We now write: " ... is calculated for each grid point by computing the difference between the maximum value and the minimum value of the different reference datasets, and this for every 3-month period (season) averaged over the 1980-2017 period." We moved this sentence as well to the end of the section.

• l. 220: specify that you calculate spatial MAE.

Indeed, we have to specify that it is the spatial MAE. We now write "The climatological means, biases, and mean absolute errors (MAE) were spatially averaged to obtain one mean value over the complete domain and each of the subdomains, respectively."
We removed the sentence "Additionally, the mean absolute error (MAE) was calculated to account for compensating errors." since it is obvious.

• l. 225: over the points of the domain

Good point, we added the sentence: "These metrics are computed over all grid points of the CAS-CORDEX domain." and now write: "averaged over the domain".

• l. 225-227: The Taylor Diagrams does not represent, but rather includes calculations of the spatial correlation between model and reference data. In your case you can eventually state that the Taylor diagrams are used to estimate spatial agreement between the climatological means of the different data-sets, by considering different metrics.

We have reformulated the sentence: "These diagrams supplement the bias analysis by visualizing in a concise way information about the spatial correlation, the centered root mean square error (RMSE) and the ratio of spatial variability (RSV) between the model and the observational dataset (Taylor, 2001)."
We have removed the sentence "In this study the Taylor diagrams represent the spatial pattern correlation between model and reference data, which is obtained by calculating correlations across the grid points of the CAS-CORDEX domain." since it doesn't add any non-trivial information to the text.

• l. 239: maybe it would be nice to indicate the different rows corresponding to annual and different values.

Here we think it is sufficiently clear when looking at the figure.

• l. 242-244: I would discuss the results of the table together for annual and seasonal values, after introducing the results of Fig. 2. In any case, you have to provide more details on the results of Table 2 for yearly values, not simply mentioning that they are reported.

As mentioned in reply to your general comments, we have put a bit more order in the structure of section 3. First the figure and table are introduced, then the annual averaged performance is discussed, followed by summarizing the seasonal averaged performance.
We added a summary of the annual results from Table 2: "The biases and MAE of the annual mean temperature are very comparable between ALARO-0 and REMO (Table 2), with small biases and MAEs that are only slightly larger than the spread of the observational data sets."

• l. 252: Very pronounced biases are also present for REMO, in particular in winter, over the north-eastern part of the domain

Indeed, we now write: "On the seasonal timescale, biases over larger areas are mainly pronounced in winter (DJF) and spring (MAM). In particular both models locally show strong biases in the north-eastern part of the domain for winter with values ranging up to 15 °C. Additionally, ALARO-0 shows strong negative biases up to -15 °C during spring in this area."
We removed the sentence: "In winter the most pronounced bias is found for REMO over the north-western part of Mongolia in the Altai mountains, resulting in a large MAE of 3.40 °C over the ESB domain." since it does not add any substantial information anymore.

• l. 253: the reader should know which areas you are referring to: the map of the subdomains should be introduced directly in the main text and not in the supplementary part.

The figure has been added to the main text of the revised manuscript.

• l. 255: as you did for ALARO, mention that the bias of REMO exceeds 10C in this case.

As mentioned in the answer to comment l. 252 we now include in the text that REMO has biases up to 15 °C in the northeast.
Additionally, we now write: "These large biases are reflected by the values in Table 2 for the northern subregions EEU, WSB and ESB for ALARO-0 and the ESB subregion for REMO."

• l. 256-258: that's it for spring? please quantify the values of the bias in the two cases, consistently with other seasons. Also, what happens over other regions?

Spring is partly treated together with winter, since the main bias in ALARO extends from winter to spring. We prefer to keep it like that since splitting the two would make the paper longer.

• l. 258-259: as above, provide estimate of biases also in summer

In fact, all the estimates are in Table 2. This is precisely why Table 2 is added; to avoid that the paper becomes (even) longer.

• l. 260: On the contrary

The correction is done.

• l. 260-261: Why do you stop providing estimates of the biases? This applies to all sections and figures.

The purpose of the text is not to provide the biases by numbers. Otherwise, the text would become even more lengthy. We only provide them to describe the main deficiencies here. For instance for summer, both models behave within the range of the observations, so we do not provide the numbers.

• l. 260-261: what about summer biases of ALARO over Mongolia and North-eastern China?

We find that they are rather small compared to the others and prefer to not extend the text too much. See our reply to your previous point.

• l. 265-268: I would move this paragraph before the previous one, at the beginning of the description of Fig. 2.

Given the structure as we proposed it in the reply to your general comments, we prefer to keep it here. Otherwise it would feature before the description of the annual performance while it describes the performance of the seasons.

• Fig. 3: How can you explain very high spatial correlation values for the entire domain when the values of correlation are lower in the case of all subdomains?

Positive and negative terms can cancel each other out. When computing the spatial correlation over the CAS domain, biases are more likely to have an opposite sign (canceling each other out) compared to the subregions. Not surprisingly, the models simulate the global climate better than the regional ones. This is what is meant with: "On the other hand, the Taylor diagrams for the subregions illustrate how scores calculated over the complete CAS-CORDEX domain can hide underlying regional trends."

• l. 275: how do you define best here? in terms of which metrics? as the points being closer to observations? be more specific. For example for REMO RSV, the model is better in summer than autumn in EEU; RSV is better in summer than Autumn over West Siberia with respect to all metrics;

for East Siberia in summer, values of RSV are closer to 1 than in autumn; for WCA, summer correlation is slightly better than in autumn; for the TIB region, both summer and spring results have better correlation than in autumn. Only for ALARO results seem to be better in autumn
than for other seasons, given the considered metrics. The description of the Taylor diagrams is not very accurate and precise throughout all the text, making the interpretation of the results very difficult for the reader. I suggest the authors to thoroughly review the comments of the Taylor
diagrams, trying in a first essay to write down all possible information and only then summarizing the results. I can understand that this is complicated given the large amount of figures, but this is absolutely necessary given the current state of the manuscript.

This was referring to the different metrics taken into account, namely spatial correlation and RMSE. The RSV is described in the sentences that follow. We reformulated the period to explain more clearly what we want to say. We consider centered RMSE <0.5 and spatial correlation > 90% here and it is made clear by the previous sentences. We would prefer not to extend the period or add the proposed detailed observations, again to make the paper not more lengthy than necessary. Most of them are already included in the second part of the sentence or in the sentences that follow. Our aim is to write down the general outcomes, not a long summation of detailed information that directly is seen from the graph.
Indeed, the RMSE for ALARO-0 is always smaller during autumn in the different subregions (and the spatial correlation is often higher), while, as mentioned in the text, this is not the case for REMO which only shows smaller RMSE values during autumn over the EEU, WCA and ESB subregions. However, REMO is closest to CRU during autumn over the full CAS-CORDEX domain.

• 277-278: not true. What do you consider as ”WELL” for normalized standard deviation? For me the good performance of ALARO in EEU in Winter amd REMO in winter over WSB are quite arguable. Also, I would not be very convinced about the goodness of the winter results for the ESB domain for ALARO as well as winter results in WCA and autumn results in TIB for both models.

Indeed, the RMSE is large in winter for some subregions, but here we are talking about the normalized standard deviation (RSV). We write “in general”. By “in general” we mean that the RSV deviates less than 0.25 from 1 in most cases, so we think the statement can be defended. We have added this to the text. If we would expand on all the subtleties it would make the paper again more lengthy.

• l. 279-280: how would the limited bias explain the higher SD values? biases could be higher but more homogeneous. Reformulate this period.

Indeed, we now write:
“During spring the cold bias in the north is limited to -5 °C for the REMO model but not for ALARO-0, which is reflected in a higher RSV for the northern regions.”

• l. 282-285: This part is not exhaustive and accurate and needs to be reformulated: first of all a similar bipolar behaviour is found for different seasons for both models, not only for summer. One good example is SON spatial biases for the entire domain for REMO and the case of both REMO and ALARO in winter. Here it would be opportune to use the fact that you have small mean biases but large MAE for supporting the conclusions on the fact that the biases are the results of compensation effects. Also, poorer performance than what? Given your conclusions about the performance of ALARO in summer, why your reasoning should work for summer and not for winter, where we also have a very well pronounced bipolar behavior of the bias?

The bipolar behaviour over the complete domain in winter was pointed out earlier in the text, when we describe Fig. 3 and Table 2. It does not make sense to repeat this here. We moved the explanation for the summer to the same paragraph.
Indeed, what we wanted to point out here is poorly formulated. We now write:
“The small mean bias during summer (JJA) for ALARO-0 over the complete domain (Table 2) is the result of averaging the warm biases in the south and the cold biases in the north (Fig. 3).” and:

"High RSVs are also observed for ALARO-0 in summer over the complete domain (Fig. 4) and this is due to the underestimation of the cold temperatures in cold regions, while warm temperatures are overestimated in regions that are characterized by warmer temperatures (Fig. 3)."
The reasoning works as well for winter. In winter a smaller RSV (RSV < 1) over the complete domain is obtained for ALARO-0 and REMO due to the warm bias in the north. This is not the case for the ESB subregion since there are also areas with cold biases that cancel the effect of the warm bias on the RSV out. In other words when colder regions are simulated colder, then the RSV is larger and when colder regions are simulated warmer, then a larger RSV is obtained.

• l. 290: specify that you are discussing spatial variability

Good point. In fact, it should be seasonal and not spatial. We now write:
"Comparing the metrics of the RCMs (Fig. 3, Fig. 4 and Table 2) shows that REMO is better in simulating the seasonal variability in temperature compared to ALARO-0, except for the autumn in all subdomains and winter in the WSB and TIB subdomain. On the other hand ALARO-0 ..."

• l. 289-293: I would reformulate this part: you basically say that REMO is better than ALARO in simulating spatial variability, except autumn and winter for some domains. Then you say that ALARO better captures spatial patterns except winter and summer over some subdomains. This sounds a bit contradictory.

Indeed, but the confusion comes from the notion of the variability. It is seasonal, see our previous reply.

• Fig. 3: specify in the text that the different Taylor diagrams have different scales.

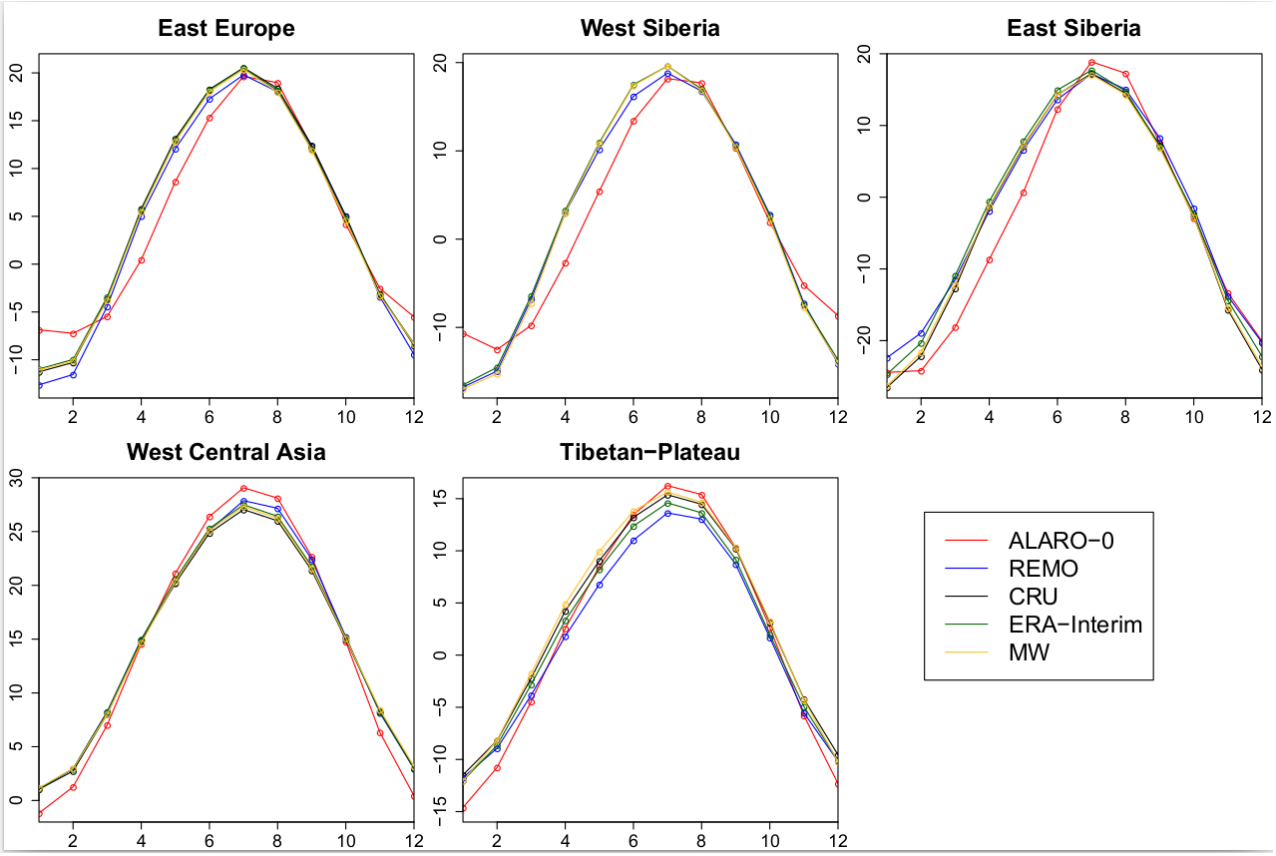
We prefer not to overload the text. It is clear when viewing the figure.

• l. 300: extremely well? it does not seem so for WCA and ESB; Also performance do not seem very good in November for ALARO over WSB.

For this sentence we based us only on Fig. 4 (now Fig. 5) and not on Table 2, Fig. 2 (now Fig. 3) or Fig. 3 (now Fig.4). The spatially averaged temperatures of the models are within the observational spread or only deviate slightly (<1 °C). With a 2 °C difference, the spatially averaged temperatures for ALARO over WSB and WCA do indeed deviate more from the observational datasets in November. We have now written: "From figure 5, it can be seen that the RCMs simulate the spatially averaged temperatures extremely well during the autumn months (months 9, 10 and 11), since they are within the observational spread or deviate slightly from the observational spread (<1 °C). The exceptions are the spatially averaged temperatures for ALARO-0 over WSB and WCA in November where the spatially averaged temperature deviates 2 °C from CRU." To make the order more consistent with the previous paragraphs we moved these sentences to the end of the paragraph.

• Fig. 4: I suggest you to split the figures in 3 figures, respectively for T2, TMIn and TMAX, since at the moment the current figure does not allow to appreciate differences in the different cases. Maybe you can introduce figures for TMIN and TMAX in the supplements.

We prefer to keep this figure in the text since it limits the length of the manuscript and all information is included. As it can be seen in the following figure for mean temperature, splitting the figure into three figures for mean, minimum and maximum temperature does not resolve in showing the differences between the datasets better. If it is not possible to see the differences visually, then the difference between the datasets is < 1 °C, which is small and not worthwhile to mention it.



• l. 326: what happens for the orographically more complex regions? What are the biases?

Indeed, that sentence is not clear. We now write:
"Annual biases of the minimum temperature over Russia in general vary mostly between -3 °C and 3 °C for REMO and between -1 °C and 5 °C for ALARO-0, with a few exceptions in the orographically complex regions, e.g. in the Stanovoy Range and Central Siberian Plateau where higher biases are found."

• l. 327: Specify that the bias of REMO over the Eastern part of the domain reaches 15C in winter.

Indeed, we now write:
"The warm biases for REMO in the Eastern part of the domain are most pronounced during winter reaching up to 15 °C. ALARO-0 also shows equally large biases, but ..."

• l. 329-330: what about all the other parts of the domain in spring and summer?

In the other parts, we think there is nothing particularly to report.

• l.345-354: Quantify values of bias and MAE.

We did not specify the values for the Table 2 within the text, since the purpose of the table is precisely to list them in a way that allows to easily overview them. We discuss the table only qualitatively. To stay consistent and to avoid that the paper becomes even longer we do not do so for Table 3.

• l. 352-352: this does not seem the case for autumn and annual values of ALARO and for summer for REMO

Indeed, but we prefer to make a general statement here. We write it now more precisely:

"The normalized Taylor diagrams in Fig. 7 confirm that, in general, the RCMs struggle to simulate the spatial pattern of minimum temperature well over the north-eastern part of the domain (ESB), while on annual level ALARO-0 is able to simulate the spatial pattern well."

• fig.6: I think you are not giving the same importance in the discussion of Fig 6 as for Fig. 3.

The reader should, at this point, be familiar with reading the Taylor diagrams. Indeed, as mentioned in reply to your previous comment, we prefer to limit us here to general statements. We prefer not to extend the paper.

• Fig. 6: one general question concerning all Taylor diagrams: do you have any clue why you generally get high values of correlation for the entire domain, when in all subdomains you obtain smaller correlations?

In fact, there is no reason why this should not be the case. Positive and negative terms can cancel each other out. When computing the spatial correlation over the CAS domain, biases are more likely to have an opposite sign (canceling each other out) compared to the subregions. Not surprisingly, the models simulate the global climate better than the regional ones.

• l. 361: not exact: REMO better also for EUU in summer and winter and in WCA for spring.

We now write: "REMO has a better centered RMSE and spatial variability during summer, except for the WCA region". It would require extra sentences to describe both RMSE and spatial variability in detail for each season since the outcomes differ quite a bit for each subregion. The manuscript is already long, so we prefer to not add this.

• l. 361: specify spatial variability

Indeed, we now write: "On the other hand, REMO has a better centered RMSE and spatial variability during summer, except for the WCA region … "

• l. 366: what happens in the Tibetan plateau? what is the magnitude of these biases?

Indeed, this paragraph can be formulated in a better way. We rewrote the paragraph including more numbers of the biases that can be seen on Fig. 7 (now Fig. 8).

• l. 365-370: not only here, but also for other figures, fix an order for the discussion of the figures (for example from top to bottom) and follow it throughout the text

As said in the reply to your general comment, we have paid attention to the structure of the text. However, this is not always possible when certain features appear for different seasons since we prefer to mention this at the same time to limit the length of the text.

• l.367-368: cold bias over northern part of the domain is present in all seasons, except DJF.

Indeed, this can be improved. We now write: "Biases in Fig. 8 and Table 4 show that for both RCMs a pronounced cold bias is present for maximum temperatures over the northern part of the domain at the annual scale and for all seasons, except for ALARO-0 in winter."

• l. 364-371: reformulate and extend all the period. Some parts of the domain are never mentioned. What happens for example in the Himalayas and the Arabian Peninsula? exhaustive description of summer is missing.

We rewrote the paragraph. Also here, we prefer to not extent the manuscript. For the Himalayas it is obvious. For the Arabian Peninsula there is nothing substantial to report. It would only make the manuscript longer.

• l. 378-379:Specify that in the case of ALARO the bias exceeds 7C.

This is substantial. We added the sentence: "Both RCMs have a cold bias over a large area in the north during spring, which is very pronounced for the ALARO-0 model in the north-east (< -15 °C), while the biases remain limited to -7°C for REMO (Fig. 8)."

• l. 376-385: you should extend the part describing the tables, discussing for example biases and MAE, giving indications on the fact that in some case biases are the result of compensation.

All the information is in the tables and the figures. We do not think that every value in the table and figure has to be repeated in the text. So we prefer to not extend the manuscript since it is already quite long.

• l. 380: not very accurate. In summer REMO is better for the Tibetan plateau in terms of MAE, while ALARO is better in winter for the EEU domain

Indeed, we adjusted this sentence.

• l. 381: how can you conclude, from the sentence before where you state that the 2 models are better in autumn, that ALARO simulates TMAX poorly in any season? actually there are 3 cases where ALARO is better than REMO: DJF in EEU and MMA and SON in TIB. Anyhow, in many cases the results of the 2 models are very similar and the MAE is very close to 1. Importantly, you again omit to specify what is good and what is bad, in terms of the given metrics.

Indeed, this sentence is incorrect, we removed the sentence: "From this we can conclude that ALARO-0 simulates the maximum temperature poorly in any season."

• l. 403: it would be better to have a map of the bias of DTR. This could help your discussion that now is too generic based only on the maps of TMIN and TMAX.

We agree that this would be useful, but the discussion is already very exhaustive and we have to put a limit somewhere. So we prefer to not extend the text.

• l. 406: the model does not restore its balance, since it is also in equilibrium in winter. Reformulate.

Indeed. We now write: "In summer the model is able to evolve to a more correct balanced state and to simulate spatial averaged minimum temperatures ..."

• l. 414-415: Where? In the other sections you started commenting the figures and then the tables. Why you change this now?

Indeed, as said in reply to your general comment, we have restructured the text.

• l. 415-420: what about annual values in table 5? what about the spread of observations?

Also here, we have to draw a limit and we cannot include all numbers in the text.

• l. 417-418: please quantify all the biases you mention.

Also here, the purpose of the table is to list the biases. If they can be found in the table there is no need to repeat them all in the text. The purpose of the text is to describe the main features.

• l. 419: are you sure the Tibetan plateau can be classified as a monsoon region?

We now write: "… over the annual cycle for both RCMs over the East Asian monsoon region, with a less notable wet bias during summer …"

• l. 418-420: which figure are you commenting now? If you are discussing Fig. 11, actually it does not seem that the bias is smaller but greater in summer, at least for ALARO.

This is now clear with the new structure of the text.

• l. 423-424: It is not totally true that the largest biases are present over extremely dry areas. One example is Northern India in Summer, presenting a remarkable dry bias despite observations are characterized by highest precipitation values.

We now write: "Some of the largest relative biases can be found in relatively dry regions …"

• l. 422-423: specify that when you talk about low precipitation you are referring to the observations.

We now write: "This is partly due to the low observed precipitation quantities in several regions …"

• l. 425: actually over Northern China in REMO the bias exceeds 5mm/month.

This sentence is wrong in this context, it was written for the Gobi desert region only. We have removed it since it does not add any value here.

• l. 433: "is also present for REMO": why also?

Indeed, it is not clear what "also" refers to. We removed it.

• l. 435-436: which model are you referring to?

Indeed. We now write: "The wet bias for REMO over ESB during spring is low …"

• l. 438 and l. 447: greater than -2mm/month

Indeed, we wanted to point out the smaller absolute value of the bias. We changed smaller than into greater than -25 mm/month here.

• l. 450-451: how can you claim that the spatial patterns are well represented by ALARO if for each subdomain, despite WCA, the considered metrics present relatively poor values (especially in terms of correlation)?

It is better in comparison to REMO.

• l. 449-452: why do you not discuss the observations?how do the different data-sets compare to each others? this is something that you should do when commenting all Taylor diagrams.

The spread between the observational datasets and the implications on the results are according to the structure of the text described in the discussion section. Again, we have to put a limit to the content of the paper, it is already long, so we do not discuss all the details of the different observational datasets.

• You can extend the discussion of Fig. 10, consistently with the other subsections.

Also here, we have to put a limit to the content of the paper, it is already long.

• l. 476-479: Please specify the cases when the bias exceeds the spread of observations. At the moment it seems like the spread of observations is smaller than the bias in almost all the domain and seasons. Is that correct? So your figures would confirm that evinced biases are more inherent to the model than to observations, over almost all the domain in all seasons. One more elegant way (and probably more useful for your goals) to determine those points where the spread of observation exceeds the bias, is by plotting a map of the bias (in the supplements) with a point in correspondence of those points where the spread is larger than the biases.

Yes, the spread of the observations is smaller than the bias for a large part of the domain. We did write this down in the next paragraph: "Figure S1 shows that for the majority of grid points the mean temperatures of ALARO-0 and REMO lie within the range of spread between the reference datasets during autumn." "During winter and spring none of the RCMs are able to reproduce temperatures that can be completely explained by the observational uncertainty over a large part of the CAS-CORDEX domain, while this is also the case for ALARO-0 during summer (Fig. 3 and Table 2)." The maps in the supplementary material give the information of those points where the spread is larger than the biases, namely the areas covered by the white color that corresponds with values < 0. Everything in red is above 0 and indicates thus that the bias between the RCM and CRU is larger than the observational spread. We chose to add a scale with multiple levels, to give additional information about where the bias of the RCMs is much larger than the spread between the observational datasets. All the information is in the current figures or can be derived from the figures.

• l. 482-484: be more precise. It is not clear what you want to express. The observational spread is significantly high over complex-orography regions and not over the entire north-eastern part of the domain.

We now write: " ...which makes the evaluation of the models less reliable over these mountainous regions."

• l. 486-488: also REMO bias exceeds the spread of the observations over large parts of Mongolia and Northern China.

Yes, but we have to put a limit to the scope of the paper.

• l. 489-490: reformulate

We now write: "For instance, the strong biases in the north-eastern part of the domain for ALARO-0 during winter and spring exceed the spread in temperatures between the different reference datasets, indicating that ..."

• l. 497-498: This is also true for EEU in autumn

No, we do not agree. Over EEU, the MAE for both RCMs is larger in autumn compared to the MAE of MW and ERA-Interim.

• l. 531: Also true over mountainous regions of north-eastern part of the domain.

Yes, but we limit the discussion here.

• l. 536-538: but also in summer, at least for ALARO

Idem as for the previous reply.

• l. 542-543: Reference needed.

We added the references Jacob et al. (2012) and Remedio et al. (2019).

• l. 542-542: what are these processes?

We do not know. We only say that we see the same features here, so it must be the same processes that are the source of it.

• l. 543-544: why should it be? reformulate. Better specify what do you mean by shift in the annual cycle?

We mean a temporal shift in the annual cycle. We now write: "The warm bias during winter and cold bias during spring in the north-eastern part of the domain for ALARO-0 are not due to a temporal shift in the annual cycle in the northern part of the domain, ..." When you briefly look at the maps it looks like there is a temporal shift because of the warm temperatures in winter and the cold temperatures in spring, but this is not the case.

• l. 546: how can you state that the bias increases when the snow-covered region expands? Have you directly analyzed snow cover in the two models? If what you affirm is true, this should be a feature of both models. However, winter biases are different in the two cases.

The zone with a warm bias becomes larger during the months when the snow cover expands. We investigated for both models and ERA5 the monthly evolution of the snow cover but we did not implement the figures here, since this is not the aim of the paper and to overcome an even longer text. ALARO-0 produces too much snow over a too large extent compared to ERA5. The snow covered regions are not the same and this makes it more complex to understand the temperature biases that are warm for the regions where both ALARO-0 and ERA5 show snow cover. Therefore we wrote: "This shows that a more complex multi-layer snow scheme might not be enough to solve the warm bias for ALARO-0 during winter. Therefore, further investigation should be done to see whether the warm bias in winter over the northern part of the domain is due to the inability of the current snow scheme to reproduce the heat conductivity of snow." Here we discuss only the biases in ALARO-0 as written in the text, since for REMO the snow covered regions agree well with ERA5. The sentence at line 546 is only an observation. We cite the analysis of Mašek (2017) that reported the same observation. We identified that this error can be traced to the interactions with the surface which manifest themselves when the snowpack is growing. More detailed research is needed to find out what exactly is going on.

• l. 545-546: are you sure about the warm bias appearing in the North during autumn when the snow appears over this region? Again, did you base these statements on some analysis of simulated snow cover? Warm bias is very limited to a very small eastern part of the northern domain in ALARO.

Yes, we did check the snow cover and it was seen that the warm bias appears over regions where both ALARO-0 and ERA5 simulate a snow cover. In October there is only snow in the most northern part of the domain. In November most of the northern subregions are having a significant amount of snow cover. Averaging over the three months reduces the warm bias present over the northern region with snow during November to only a small area with warm bias in the northeast. The cold bias in the northeast can be explained by the fact that this is a region that is covered by snow for ALARO-0 but not for ERA5 during autumn.

• l. 548-549: I would argue against your conclusions, based on the fact that also REMO is characterized by a very warm bias in winter over the North-eastern part of the domain. I do not think that based on your analyses you can raise strong conclusions on the driver of the bias over the northern part of the domain in both models. For sure I would not state that REMO does not encounter the same problem.

Perhaps REMO also has underlying problems but more research should be done to investigate this and to know the cause. And if it has, it does not manifest itself as strongly as ALARO-0.
The warm bias for REMO does rather occur for regions where there is no snow (or not enough snow) simulated by REMO, while there is a (larger) snowpack present for ERA5. The opposite is true for ALARO-0, there is more snow simulated than given by ERA5, so as we mentioned in the text, the biases cannot be explained by the same feature.

• l. 570: wouldn't Eastern Europe be more consistent with your discussion than Western Russia?

No, since we aim here to exclude some East-European countries such as Poland, Lithuania, Latvia, Estonia… where the bias in minimum or maximum temperature is larger than the spread (now Fig. S2 and S3). When investigating Fig. 4 (now Fig. 5) it is seen that the averaged values of the EEU subregion are not completely within the observational range for minimum temperature, so we cannot change it to EEU.

• l. 571: Can you better specify what means acceptable? I would emphasize that the MAE in this case is smaller than the MAE between ERAInterim and CRU.

This is literally what we wrote in the second part of this sentence: "... since the MAE between ERA-Interim and CRU is larger (Table 3 and 4)." To make it even more clear we now write: "...since the MAE between ERA-Interim and CRU is larger than the MAE between REMO and CRU (Table 3 and 4)."

• l. 575-577: How can you state this without a map of the bias in DTR? can you be sure that for all the points of the domain RCMs produce a smaller DTR in all seasons?

Indeed, but we do not intend to make a mathematical statement here. We now write: "Both RCMs generally produce a smaller daily temperature range ..."