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

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.
- 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.

- 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.

- 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.

- 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.

- 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?

- 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.

- 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.

- 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?

- 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.

Beside these major concerns there are a series of minor, but still relevant issues that need to be addressed by the authors.

Minor Comments

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

- 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.

- line 56: Same as above. Replace validation with evaluation everywhere 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).

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

- 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.

- 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.

- line 80: complemented by

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

- 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.

- 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.

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

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

- 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.

- 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.

- 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.

- 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).

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

- 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.

- 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.

- 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?

- 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.

- 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.

- 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.

- lines 207-209: reformulate this period.

- line 219: seasonal means of

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

- 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.

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

- 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.

- 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.

- 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?

- 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.

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

- 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.

- 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.

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

- 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.

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

- 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.

- 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?

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

- 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.

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

- 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).

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

- line 315: less good than what?

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

- 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.

- end of line 326: the cold bias

- 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.

- 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.

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

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

- lines 364-366: This part should be moved 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?.

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

- 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.

- 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.

-

- line 393: systematically

- 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.

- 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.

- line 404: Why hence?

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

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

- 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.

- 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.

- 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.

- 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.

- 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.

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

- 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.

- 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.

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

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

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

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

- 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?

- 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.

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

- 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?

- 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.

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

- 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.

- line 520: "that that". Correct.

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

- 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?

- 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?

- 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?

- 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.

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

- 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).

- 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?