Reply to

1st Reviewer

Russo, E., Kirchner, I., Pfahl, S., Schaap, M., and Cubasch, U.: Sensitivity studies with the Regional Climate Model COSMO-CLM 5.0 over the CORDEX Central Asia Domain, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-22.

Dear reviewer,

Thank you very much for your effort in reviewing our paper.

Below we go point by point through your technical corrections, presented in *italic*, detailing how we dealt with your concerns reported in *Bold*.

Thank you.

General Comments

- *T*he level of the manuscript is quite poor; especially the Introduction is too "educational" and seems more similar to a technical tutorial for PhD students than to a scientific paper for experts. Many concepts are explained in details, but they are well known by the scientific community working with RCMs, and could be replaced by proper references. On the other side, a detailed synthesis of the state of the art is completely missing, especially for what concerns previous sensitivity studies performed with CCLM or other regional models. There are specific works by Bellprat et al. (2011, 2012) or Bucchignani et al. (2016) that have been referenced, but were not mentioned. In this view, I suggest the following papers to be analyzed and referenced:

  Avgoustoglou E., Voudouri A., Khain P., Grazzini F., Bettems J.M.: Design and Evaluation of Sensitivity Tests of COSMO model over the Mediterranean area. Perspectives on Atmospheric Sciences, 1:49-55.

  Voudouri A., Khain P., Carmona I., Avgoustoglou E., Kaufmann P., Grazzini F., Bettems J.M. (2018): Optimization of high resolution COSMO model performance over Switzerland and Northern Italy, Atmospheric Research, 213:70-85.

We agree with the referee and think that the introduction could

be improved and made more easily readable. Also following the suggestion of the second referee, sensibly shortening the part of the introduction on the description of vulnerability of Central Asia to the effects of climate change should help in this sense. At the same time, we will try to expand the part of the introduction on the state of the art of model performance evaluation and model calibration. For this we will consider the suggested references, together with additional ones. We will additionally try to express concepts well known in the community in a shorter form, when possible.

- *M*y major concern is however related with robustness and significance of the results. A key aspect is that none of the simulations represents correctly the observed fields used as reference. Consequently, there is no value in analyzing the relative performance of the simulations, simply because all do in a terrible bad way. A temperature bias larger than 15 °C or a precipitation bias larger than 200 % is not acceptable.

  A key question is about the reason of this shortcoming. Is it due to errors in the reference CCLM model configuration? a key parameter is certainly the time step adopted (dt), whose value is not specified in the manuscript, but it needs particular care. Alternatively, is it due to errors in the boundary conditions? The authors decided to use NCEP2, but I personally would prefer ERA-Interim, which are characterized by higher resolution, so reducing the resolution jump (other critical aspect).

The concerns raised by the referee are very important. Despite we think that we partly tried to take care of them already in the manuscript, at least partially in the experiments design, we recognize that we were not very careful in the treatment of some points, not clearly specifying them in the text. Trying to answer to the referee comment, in a first place we want to highlight the fact that the paper, given the large amount of tests conducted, is very important in order to understand whether the evinced biases are characteristic of the model itself and if they can be reduced by properly configuring the model and to which degree. For this, in our experiments we tried to be careful in as many points as possible as to isolate different sources of uncertainties. For this purpose, we performed additional simulations driven by

ERA-Interim to test the effect of different boundaries. The results of these simulations are not particularly different than the ones driven by NCEP2, presenting a very similar bias with the considered observations, for all variables (Fig. 1). We realized that we did not carefully discussed this point in the paper and we will better highlight it in the new version of the manuscript. With this additional simulations we aimed to show that the given biases are not due to the effect of considered boundary data. Or at least that, in our case, the use of mainly employed higher resolution reanalyses such as ERAInterim as boundaries, does not affect significantly model performances. Additionally, the improvement of the results when using NCEP2 or ERAInterim, with the finally proposed configuration (experiment q), is also very similar in the two cases (Table 4 of the original manuscript). We want to claim the fact that we decided, very consciously, to use NCEP2 instead of normally employed ERAInterim, for a clear reason: to try to reproduce the resolution jump (mentioned by the referee) that we will have using GCMs for the CORDEX-CORE simulations. In fact, the plan within the COSMO-CLM community is to use, for the CORDEX-CORE simulations, 3 GCMs: MPI-ESM, HadGEM and NorESM. Their resolution is respectively around 210km, 210x140km and 270x210 km. In this sense, also considering the results of the mentioned simulations, we think that our choice of using NCEP2 instead of ERAInterim is more than justifiable. We realized that we were not accurate enough on this point in the text. In the final version of the manuscript we will include information on the three models and (as supplementary material) the picture of the bias of the ERA-Interim driven simulation, together with related information. Other conducted simulations that we performed and that could help in isolate the reasons for the model bias are the two additional simulations that we performed with diffetrent soil layers and the CCLM multi-layer snow model (Fig. 2). These results are important because they show that it is not possible to improve model performance in terms of winter temperature over Siberia following previously suggested hypotheses (the snow model produces even warmer biases over the area in winter), and the given bias is very likely due to the model formulation itself. This is very important because it highlights the necessity to put more efforts in the improvement of the simulation of snow processes and permafrost in COSMO-CLM. We

will provide the new figures of the bias of winter temperatures for these simulations in the supplementary part of the final version of the paper. The only model uncertainty factor that we did not consider in our former version of the manuscript is the different time step. For tackling this issue, following the referee comment, we now conducted a new simulation with a different timestep of 120s instead of 150s. The biases against observations calculated for this new simulation are reported in Fig. 3 (in this case only CRU is used for T2M and DTR, while GPCC is considered for PRE). As you can see, using a different time-step the results do not change significantly. This confirms even more that the given biases are characteristic of the model and do not depend upon the referee suggested sources of uncertainty. We will add this figure as supplementary material to the new version of the paper, and will better discuss it in the final manuscript.

Beside these considerations, the most important thing that must be considered in order to properly address the referee comment concerning the validity and significance of the results is the comparison and proper discussion of the different observational datasets. It is certain that the evinced model biases are quite remarkable in some cases, in particular during winter for temperature. For precipitation, the biases exceed 100% over large parts of the domain, but not 200% as suggested by the referee. This is normal for many RCM studies, especially for the Tibetan Plateau, as already indicated in the former version of the manuscript. Biases of 100% in precipitation are evident in most of the CORDEX simulations, for different domains and models (Kotlarski et al. 2014, Russo & Cubasch 2016, Park2013, Martynov2013). In any case, these biases, supported by the already mentioned analyses, certainly confirm the importance of trying to investigate different model configurations in order to improve model performance over the area and to determine to which degree this is possible. We think that the paper definitely gives a significant contribute in this sense. On top of that, we agree with both the referees that the different observational datasets need more attention in the manuscript and their main drawbacks and differences need to be properly discussed. For this purpose we propose to include in the final version of the manuscript Fig. 4 of the current response, showing the spread of the different observational datasets, for each variable. This will

contribute to support the validity of the presented results. In fact, the spread of the different observations is larger in correspondence of those points characterized by particularly complex topography, for which model biases seem to be more remarkable, exceeding 15 °C in the case of temperature. This suggests that the particularly high biases evident in the model are hard to be quantified for these points. Additionally, if we now consider the new figures of temperature bias (Fig.5, together with the corresponding Fig. 6 for the DTR bias), that we drew considering the suggestion of the second referee to use a colorbar with fewer breaks, we can see that the very high temperature biases exceeding in some cases 15 °C are mainly relative to the UDEL dataset and, in general, particularly large biases are limited to a few points characterized by particularly high topography, where the gridded datasets are less reliable. If we consider the CRU dataset, that is one of the most employed dataset for the area, for evaluating the results of RCMs, we see that the values of the bias rarely exceed (are below) 10 (-10) °C, for really few points. Beside these points, still some remarkable biases are present but we think that these are well within the ranges of model simulations produced in CORDEX. For example, for the East Asian CORDEX domain the studies of Wang et al. 2013 and Bucchignani et al. 2014 showed that the CCLM over the Eastern Part of the Tibetian Plateau has a cold bias in winter, much lower than -6 C. The other area for which largest biases are present for temperature in our simulation is Western Siberia. For this region, one of the only references is the one of Ozturk et al. 2015. Their results are part of already published CORDEX results. From these, using REGCM4 with a resolution of ˜ 50 Km, they obtained a model climatological bias for winter temperature exceeding 8 °C over the area, when using CRU as a benchmark for the evaluation of their results. This is very similar to our case. Following this discussion, we propose to introduce in the new version of the paper the figure with the spread of observational datasets, together with an appropriate discussion on the possible reasons for their differences, and the new plots for the map of the bias of the reference simulation. One important thing that also needs to be highlighted in the text is certainly the fact that, beside high biases between model and observations are evident for some points, mainly where the observations are less reliable, the pattern of these biases is in general very similar among

different observational datasets and well within the range of other CORDEX results. We think that this information is certainly required in order to support the significance and robustness of the results and we will include it in the new version of the manuscript.

Always in the context of significance of the results, for the rest of the analyses of climatological values, we proceeded separately considering the differences among different observational datasets. In fact, we calculated for each dataset independently the MAE and the relative skill score with respect to the reference simulation, with the aim of choosing the best performing configuration that shows the same positive effect among all the different observational datasets. We will try to express this more carefully in the paper, where we think it was not carefully stated before. Additionally, we propose to introduce Fig. 7 as a supplement to the final version of the paper. This is obtained in a similar way as in Bellprat et al. 2012, but only considering the climatological mean of each month over the considered simulation period. Basically, we selected a reference dataset for each variable (CRU for T 2M, GPCC for PRE and CRU for DTR) and then we calculated the MAE in each case, weighting the absolute error over each point by the sum of two uncertainty sources, one given by the standard deviation of the observational datasets and the other by a standard deviation representative of considered model errors, calculated among the climatological means of the reference simulation, the one with a different time step and the reference simulation driven by ERAInterim. This is the formula we used for the new plot:

$$MAE_w = \frac{1}{TJI} \sum_{t}^{T} \sum_{j}^{J} \sum_{i}^{I} \frac{|m_{t,j,i} - o_{t,j,i}|}{1 + \sigma_{o_{t,j,i}} + \sigma_{iv_{t,j,i}}} \qquad (1)$$

where t represents the considered month and j and i are the spatial indices of the points of the domain. The two terms m and o are, respectively, the model and the observational monthly means calculated for each month of the domain. σ_o represents the mentioned observational uncertainty and σ_o the one of the model. The 1 in the denominator has been added in order to avoid infinitive numbers when the uncertainties are considerably close to 0. In this

way, points of the domain with higher uncertainties will receive a lower weight in the computation of the MAE. As it is possible to see from Fig. 7 of this document, the results of the SS using the weighted MAE are approximately the same as in the original figure 6 of the manuscript. Some differences are present in some isolated cases, mainly for precipitation, due to high uncertainties over some area. Nevertheless, the new analyses overall confirm the results of the plot of the total skill score of the paper: simulation q has the best performance for the domain, for all variables, and most of all coherently among different observational datasets and also considering different model uncertainties. All the analyses, considering different uncertainty sources in a different way, give the same results.

In conclusion, in the light of the presented discussion, better considering the mentioned points and different sources of uncertainty, we can affirm that our results are very important and most of all significant for the improvement of model performance over the area. Generally, the model, excepted some isolated points for which the bias against observations is not really quantifiable due to the large spread of the observations, does not perform particularly worse than other models normally considered in CORDEX. For sure there is some limited number of points for which climatological biases are very high, but these seem to be related to observation uncertainty. An appropriate discussion on the sources of these biases, most likely related to the complex topography of some region, will be included in the final version of the paper.

- *T*he paper does not investigate the origins of these strong biases. Section 3 contains only a rough (boring, in some places) description of the figures, but no significant insights are provided. Some considerations are provided in the Conclusions, but this is not the right place, since conclusions should be focused on the benefit of sensitivity runs with respect to the reference one.

We agree with the referee that the results part is in some cases boring and we will revise the text to make it more easily readable. For this purpose, we propose to place the figure on the investigation of seasonal biases to the supplement part, replacing it with

the referee proposed analysis of regional behavior of the different simulations. We think that in this way some not particularly relevant part of the text could be summarized in a simpler way, while focusing on single regions could help supporting the given conclusions on best model configuration. Concerning the current structure of the paper, we actually want to keep the results part as a description of the figures (the results) as we conceived it in a first place. On the other hand, we want to keep the part on the discussion of the results in the conclusion, but, following the referee comment, changing the title into a more appropriate "Discussion and Conclusion" section.

- *I*n Sec. 2.3, you have properly defined some subdomains, but then they are used only for the analysis of variance. Instead, the results in terms of MAE (presented in Figs.6 and 7) are averaged over the whole domain, which is too big and includes very different climate areas. I reccomend that further investigations in terms of MAE be performed considering the single subdomains.

We agree with the referee that further investigation in terms of MAE performed considering single sub-domains is required. This in fact could help to have a more proper idea of how model results change for different areas characterized by different climate conditions, contributing to the determination of an optimal model configuration and to better discriminate reasons of possible shortcomings. Therefore, we propose to introduce Fig. 8 of the current document, representing the SS of the MAE calculated over single sub-domains, in the final version of the paper, together with an appropriate discussion of the results. The proposed figure is obtained in the same way as in the case of the entire domain: the calculation of the MAE is conducted separately for the different observational datasets. For visualization reasons, we propose to plot the results of the analysis per sub-region for a single observational dataset for each variable (CRU for T 2M and DTR and GPCC for PRE), with a point for each region for which the given configuration produces the same model response among the different observational datasets. Fig. 9 shows the SS of the MAE calculated for different sub-regions for all the considered observational datasetes. We propose to include this figure in the

supplement part of the manuscript. At the same time we propose to also include as supplement to the final version of the paper, Fig. 10 of the current document, presenting the same analyses per sub-region but using the weighted MAE. As we can see from Fig. 8, Fig. 9 and Fig. 10, beside some exceptions, the results have a similar behavior for all different cases, with the same conclusions that can be drawn. These plots help because they allow to investigate model behavior for single regions, as already said. In particular, they allow to see that a complete improvement of model results over all the sub-regions is not completely achievable. As discussed in the introduction of our paper, one has always to be aware of the fact that calibrating the model could lead to better results, but this might also be the result of compensating errors. Reinforcing these thoughts, we think that with the proposed optimal configuration q the model improves in large part of the cases, for all variables. These results highlight again the advantages of using configuration q for the region. The newly proposed analyses also allow to see that in some cases model improvements almost reach 35% with respect to the reference simulation. This, once again, adds significance to the proposed results. In the final version of the manuscript we will change the results part as already proposed, substituting the plot of seasonal results with the one of the analyses for sub regions. The text will be changed accordingly to the new introduced results, trying to make it more easily readable.

- *F*inally, the differences among observational datasets are not discussed and the possible reasons for these discrepancies are not investigated,

We agree with the referee. As already stated above, we aim to introduce in the new version of the manuscript a proper discussion on the different observational datasets, their differences and the possible reasons for them.

Specific Comments

- *P*ag 2, Lines 11-18: "Among the...at once". This paragraph contains too many geographical and economical details and in my opinion does not fit well in the Introduction.

We understand this issue, raised by both the referees. We realize that this part should be significantly shortened, being only secondary to the objectives of the paper. We think that this would also help making the introduction more easily readable.

- Pag 2, Lines 25-33 and Pag 2 Lines 1-9: "The countries...due to climate warming". These paragraphs are rather an analysis of the implications of climate changes on this area, and in my opinion do not fit well the aim of the work. They should be significantly shortened.

We agree. We will shorten this part as proposed in the previous answer.

- Pag2, lines 9-11: "All the reported...strategies". This sentence is a repetition of concepts already expressed.

We agree. This part is repetitive and we will delete it from the final version of the manuscript.

- Pag 2, Line 14: "Assessing...Evaluation". This definition is repetitive and can be removed

We will remove this line accordingly to the referee comment.

- Pag 2, Line 17: "Model Evaluation...development". This sentence is prosaic and can be removed.

We agree with the referee and will correct the text accordingly.

- Pag 3, Line 27: "A series...simulation". This sentence is prosaic and can be removed.

Same as above.

- *P*ag 4, Lines 11-14: "In the light...are presented". From this sentence, I do not see a relationship between your sensitivity and the CORDEX-CORE activities. Please explain better this relation and, at the same time, explain what CORDEX-CORE is.

CORDEX-CORE stands for CORDEX - Coordinated Output for Regional Evaluations (CORE). This is the next phase of the CORDEX initiative, designed in the light of the upcoming IPCC report, with the objective of coordinating a set of high resolution climate projections for different regional domains, including Central Asia. In this perspective, our work represents the first step for the production of climate projections for the Central Asia domain using COSMO-CLM, evaluating general model performances, isolating the effects of different uncertainty sources on model results and determining an optimal model configuration for a region region for which almost no reference exists. Following the referee comment we realized that we probably did not specify very well this information in the former version of the manuscript. Consequently, we propose to extend the relative part of the text accordingly.

- "This study...domain". This concept has already been expressed in the Introduction. Please put it only once.

We will remove this repetitive part of the text, following the referee comment.

- *F*igure 1: It is preferable to show the domain using the geographical coordinates, since the reader is generally not interested in the rotated coordinates (being rotated coordinates used only internally for COSMO-CLM calculations)

We agree. We propose a new version of the map of the domain topography, presented in Fig. 11 of the current document, in geographical coordinates. The figure caption will be modified accordingly.

- *P*ag 5, lines 11-13: "The model configuration...en)". These details are not necessary, epsecially because readers are generally not authorized to download the model configuration from the website of CLM Community. Please add more details about model configuration in Table 2.

We agree with the referee that the description of the model configuration introduced in Table 2 of the previous manuscript version needs to be extended. Also, the link to the CLM-Community webpage could be removed since, as suggested by the referee, not all users could access the given configuration. Proposing to modify the text accordingly, we still want to mention the fact that we use as a benchmark for our reference simulation, the configuration of COSMO-CLM for another CORDEX domain, but covering a large part of Central Asia. This follows the main guidelines of the CLM-community for the model configuration.

- *T*able 2 (caption): The general description of model setup of the reference simulation is very poor. It contains details that have already been explained in the text (e.g. spatial resolution, domain extent). Btw, the domain extent must be expressed in terms of max/min longitude/latitude and not in terms of number of grid points. Some important details of the model setup are missing. For example, in Table 1 you write that in b configuration Tegen aerosol is used, but what is the aerosol used in a? I guess the default Tanre, but you have to specify it. Similarly, for albedo: what is the default one? I guess albedo as function of soil type.

We agree with the referee and will modify table 2 accordingly to his comment. Information about the time step of the reference simulation (150s), the Aerosol, for which we used TEGEN as default, and the albedo, as a function of the soil type, will be included, together with information on the domain extent expressed as min and max lon and lat.

- *P*ag 5, lines 19-20: "since their...simulations". This is not a good reason to employ NCEP reanalysis as driving data. Generally, data at higher resolution are preferred. Btw, what is the resolution of GCM normally employed in CORDEX simulations

As already stated above, we were completely aware of the decision taken using NCEP2 reanalyses instead of ERAInterim, with the main goal of simulating a jump in resolution more similar to the one using the GCMs mentioned above. For this, we consider our choice more than valid. NCEP2 are still considered a valuable reanalysis dataset, that has been used in a large variety of studies. Beside that, we also conducted similar analyses with ERAInterim to have an estimate of the effect of using higher resolution drivers on the results and how they change in the different cases. We demonstrated that the effect of the two different datasets on the simulation of climatological monthly means for the considered period is almost the same. We will highlight this point better in the new version of the manuscript.

- Sec 2.2: it is not clear if the original resolution of datasets CRU, UDEL, MERRA GPCC is 0.5°or if they have been interpolated on a common grid with common 0.5°resolution.

The resolution of the mentioned datasets is all 0.5°. No interpolation was needed. We will try to make it clearer in the new version of the manuscript.

- Pag 6, lines 19-20: "The climate...interpolation". This technical detail (usage of CDO) is not necessary and can be removed.

CDO is an important tool for the postprocessing of climate data, freely available. It is a personal decision, but also following the work of other papers, given its importance, we think that it deserves to be referenced in the text.

- Pag 7, line 25: "It is not specified if variances (observed and simulated) are evaluated starting from monthly values.

We acknowledge the fact that we have not been very clear in this sense in the previous version of the paper. We now propose to modify the final version of the paper better specifying that the variance is evaluated starting from monthly values.

- *P*ag 8, line 12: A bias of 15°or larger is not acceptable.

Again, this seems to be a problem related more to the reliability of the gridded observational datasets over some points rather than to the model itself.

- *P*ag 8, line 12-21: In this paragraph you are commenting Figure 3, which is related to simulation a, so it is not wise to comment here also the simulations SOIL and SNOW.

We agree. We will try to introduce the results of simulation **SNOW** and **SOIL** in an additional separated subsection of the results part.

- *P*ag 9, line 1: Why do you claim that this sentence is "interestingly"?

We think that "interestingly" in this case could be deleted.

- *P*ag 10, line 10: Why in this case analyses focused on a single observational dataset?

In the figure of the variance ratio we showed the results just for a single observational dataset for each variable, simply for graphical reasons. Nevertheless, the same analyses were conducted for the different observational datasets, and considered when discussing the uncertainties in the estimation of simulated variance. Realizing that this was not clearly specified in the text, we will modify this part accordingly.

- *P*ag 11, lines 17-18: "For the experiments... experiment q". This sentence is just a repetition of the sentence at lines 13-15. Please combine them.

We will try to merge the two sentences together as suggested by the referee.

- *P*ag 11, line 24-25: "This indicates...driving dataset". This sentence is very strong and must be supported by results that are more robust.

The few numbers shown in Table 4 are not sufficient. Moreover, you should add in Table 4 the improvements achieved when using NCEP2, in order to have a quantitative comparison.

We agree that the given sentence is too strong. Nevertheless, it is true that the 2 conducted ERAInterim-driven simulations present similar climatological values to the NCEP2-driven ones, in both cases. Even though it is not possible to draw a general conclusion on the effects of the boundaries on COSMO-CLM for the region and the given resolution, these results allow to justify the use of NCEP2 instead of commonly employed ERAInterim for the purposes of our research. We will modify the corresponding part of the text in the final version of the paper, being more careful on the conclusion we can draw from our simulations.

- *P*ag 11, lines 32-33: "Values closer... observations". This statements are obvious and can be removed.

We agree and will modify the text accordingly.

Minor Comments

- *P*ag 6, line 23: You have already explained that the reference configuration is the a. Please remove "(a,Tab.2)".

We agree and will modify the text accordingly.

- *P*ag 7, line 6: Do you mean Tab.3 (instead of Tab.4)? Otherwise, Tab.3 is never referenced.

We acknowledge the error. We referred to Tab.3. We will correct it in the final version of the manuscript.

- *T*itle of Fig. 6: If SS is defined according with equation (1), why did you add (%) next to SS?

Actually we propose to express SS in %

- *l*ines 16-17: In the title you use NCEP, in the text NCEP2, please use always the same acronym.

We will correct the error in the final version of the manuscript.

Below we propose some additional bibliography that we will provide in the revised version of the manuscript, if not already considered, following the referee comments and the proposed discussion.

References

[Ozturk et al. (2012)] . Ozturk, T. and Altinsoy, H. and Türke, M. and Kurnaz, M.L., 2012. *Simulation of temperature and precipitation climatology for the Central Asia CORDEX domain using RegCM 4.0*, Climate research, 52, 63–76.

[Bucchignani et al. (2014)] . Bucchignani, E. and Montesarchio, M. and Cattaneo, L. and Manzi, M. P. and Mercogliano, P., 2014. *Regional climate modeling over China with COSMO-CLM: Performance assessment and climate projections*, Journal of Geophysical Research: Atmospheres, 119, 121–151.

[Wang et al. (2013)] . Wang, D. and Menz, C. and Simon, T. and Simmer, C. and Ohlwein, C., 2013. *Regional dynamical downscaling with CCLM over East Asia*, Meteorology and Atmospheric Physics, 121, 39–53.

[Kotlarski et al. (2014)] . Kotlarski, S. and Keuler, K. and Christensen, O.B. and Colette, A. and Déqué, M. and Gobiet, A. and Goergen, K. and Jacob, D. and Lüthi, D. and Van Meijgaard, E. and others, 2014. *Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble*, Geoscientific Model Development, 7, 1297–1333.

[Park et al. (2013)] . Park, J.H. and Oh, S.G. and Suh, M.S., 2013. *Impacts of boundary conditions on the precipitation simulation of RegCM4 in the CORDEX East Asia domain*, Journal of Geophysical Research: Atmospheres, 118, 1652–1667.

[Martynov et al. (2013)] . Martynov, A. and Laprise, R. and Sushama, L. and Winger, K. and Šeparović, L. and Dugas, B., 2013. *Reanalysis-driven climate simulation over CORDEX North America domain using the Canadian Regional Climate Model, version 5: model performance evaluation*, Climate dynamics, 41, 2973–3005.

[Russo et al. (2016)] . Russo, E. and Cubasch, U., 2016. *Mid-to-late Holocene temperature evolution and atmospheric dynamics over Europe in regional model simulations*, Climate of the Past, 12, 1645–1662.

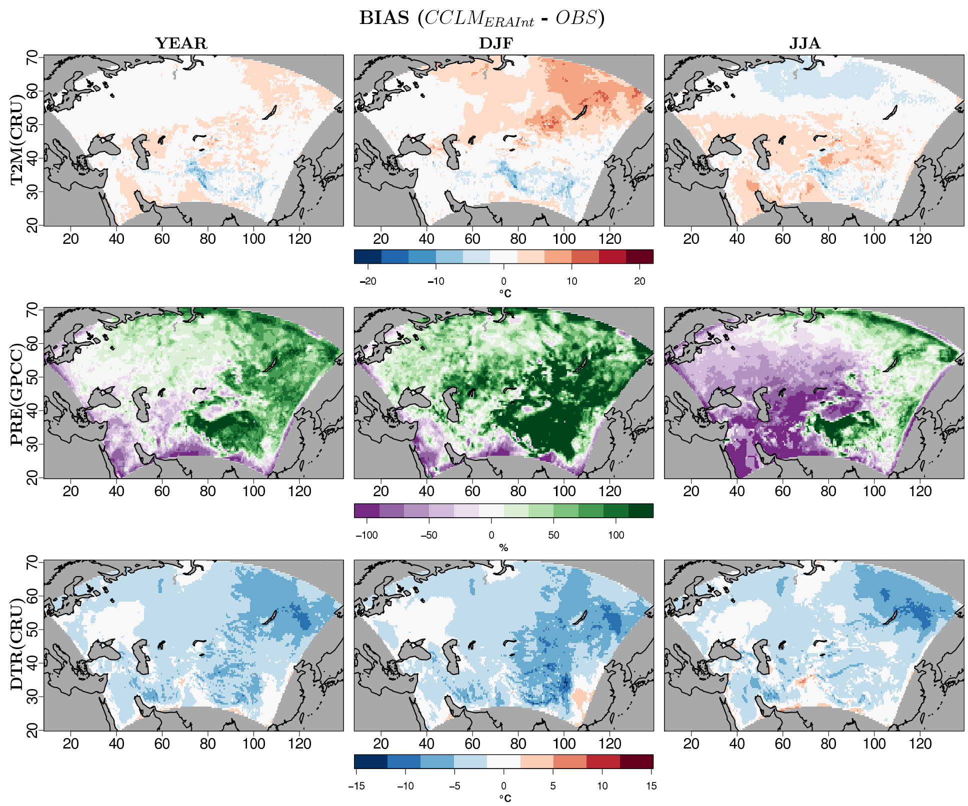

Figure 1: Mean bias of annual mean (left), winter mean (middle) and summer mean (right) near surface temperature (T2M, °C, top panel), precipitation (PRE, %, middle panel) and diurnal temperature range (DTR, °C, bottom panel) of the reference COSMO-CLM configuration (a), driven by ERAInterim reanalysis, with respect to 3 considered observational datasets (from top to bottom: CRU, UDEL and MERRA2), for the period 1995-2005.

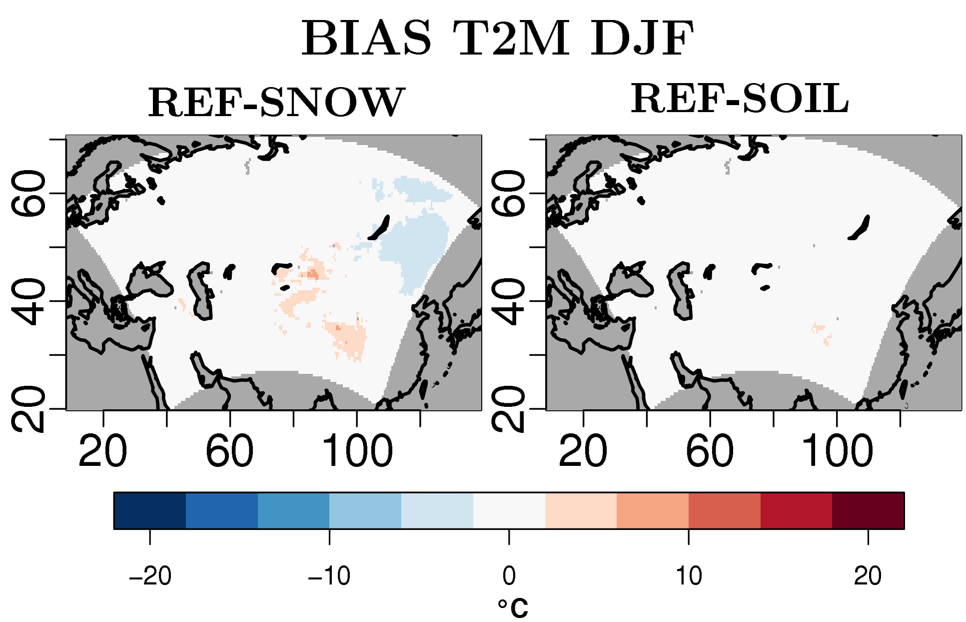

Figure 2: Mean bias of near surface temperature (T2M, °C) winter values of the reference simulation **a** and the simulation **SNOW** (*left*) and **SOIL** (*right*) , all driven by NCEP2, calculated over the period 2006-2015.

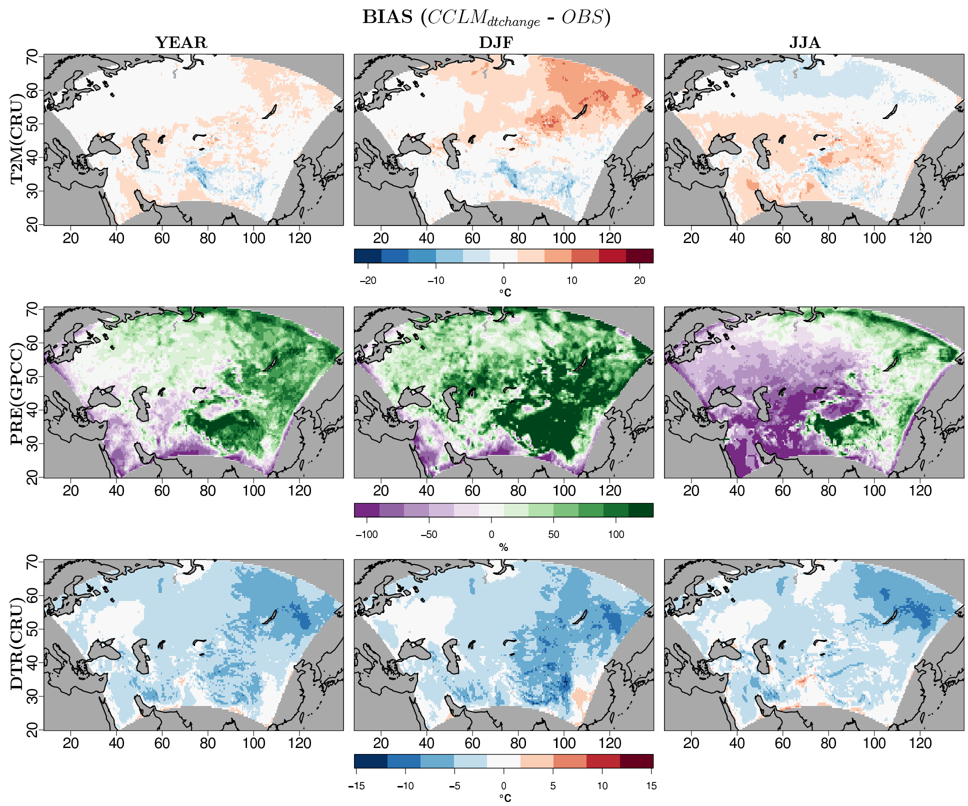

Figure 3: Maps of bias calculated for the NCEP2-driven simulation with the reference configuration but different time step, against different observational datasets for the 3 considered variables. Every row presents the bias calculated for the different considered variables, from top to bottom, respectively, near surface temperature (T2M, °C, CRU), precipitation (PRE, %, GPCC) and diurnal temperature range (DTR, °C, CRU). From left to right, values of the biases for yearly, winter and summer mean climatologies over the period 1996-2005 are presented.

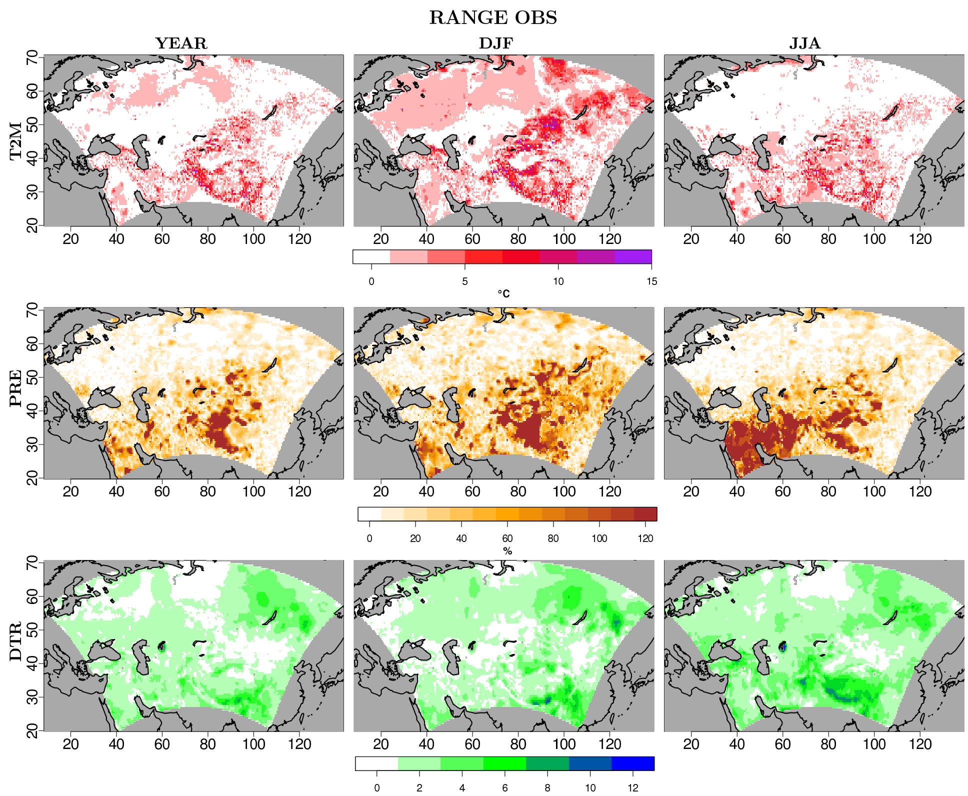

Figure 4: Maps of the spread calculated among different observational datasets for each considered variable, for the annual mean (left), winter (middle) mean and summer mean (right). From top to bottom, the values for near surface temperature (T2M, °C), precipitation (PRE, %) and diurnal temperature range (DTR, °C, bottom panel) are respectively represented.

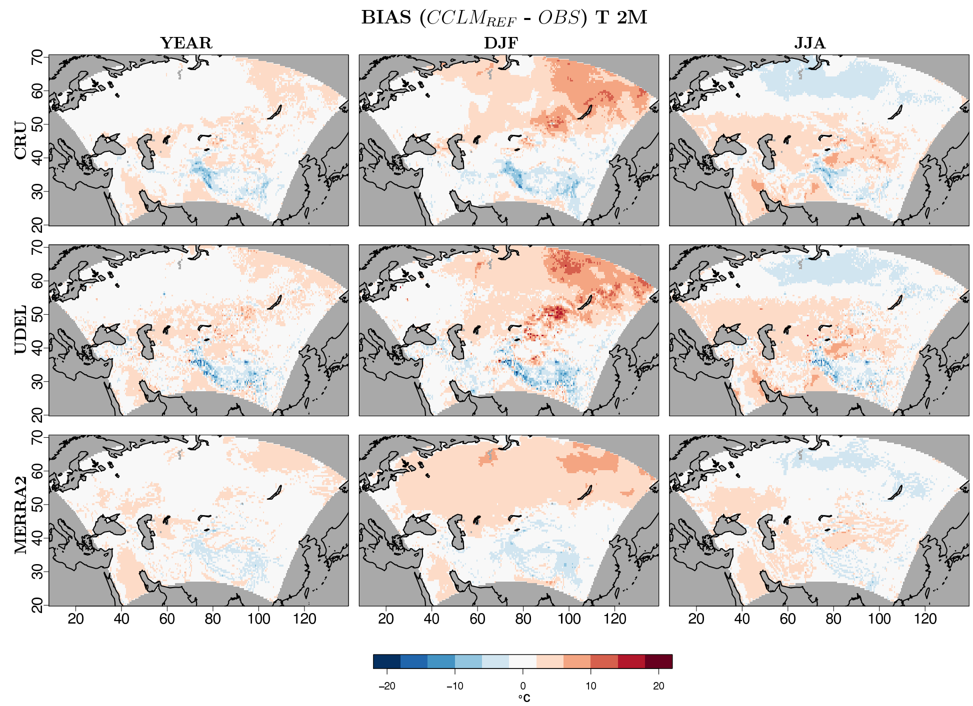

Figure 5: Mean bias of annual mean (*left*), winter mean (*middle*) and summer mean (*right*) near surface temperature (T2M, °C), of the reference COSMO-CLM simulation (a) with respect to three observational datasets (from top to bottom: CRU, UDEL and MERRA2), for the period 1995-2005.

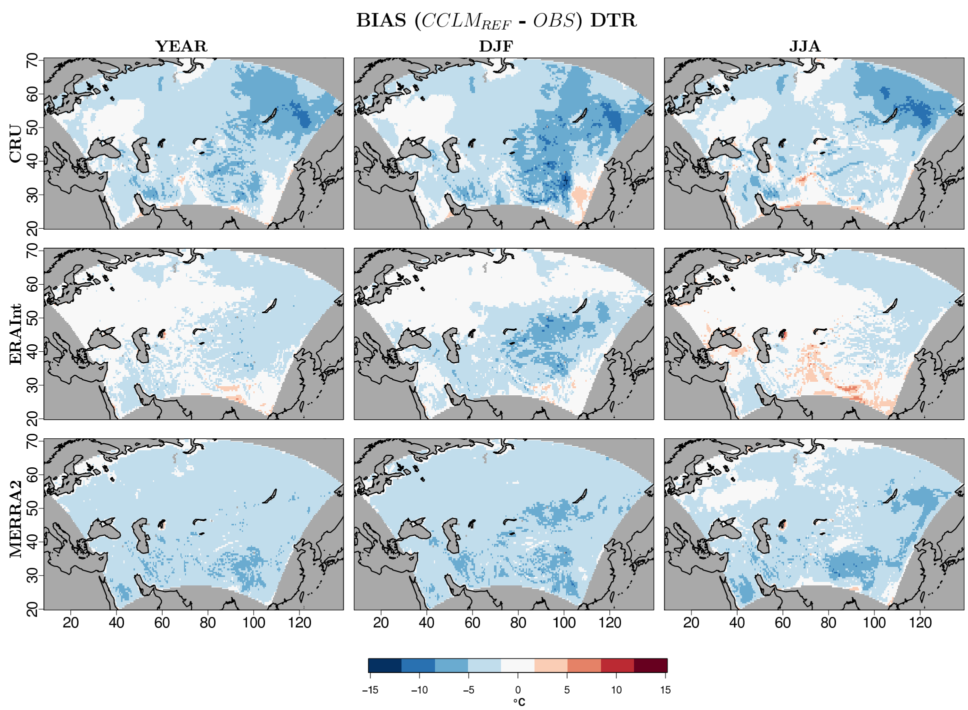

Figure 6: Mean bias of annual mean (*left*), winter mean (*middle*) and summer mean (*right*) Diurnal Temperature Range (DTR, °C), of the reference COSMO-CLM simulation (a) with respect to three observational datasets (from top to bottom: CRU, MERRA2 and ERAInterim), for the period 1995-2005.

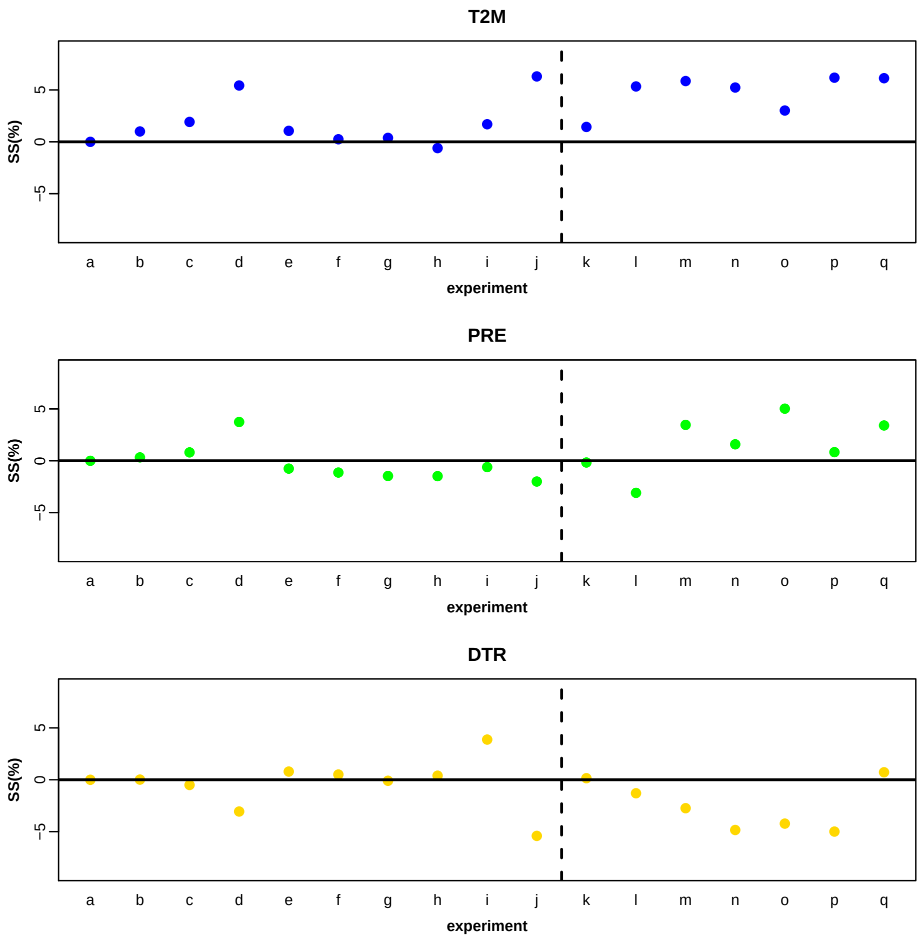

Figure 7: Skill Score (SS) derived from the weighted MAE (MAE_w, Eq. 1) calculated over the monthly climatological values of the seasonal cycle of different COSMO-CLM simulations and observational datasets. From top to bottom, the SS for each variable is displayed. The dotted vertical black line divides the simulations with the same configuration of the reference simulation plus a single change in the model setup (on the left) and the ones obtained through the combinations of the previous ones (on the right). Positive (negative) values indicate better (worse) performance of the considered simulations compared to the reference one.

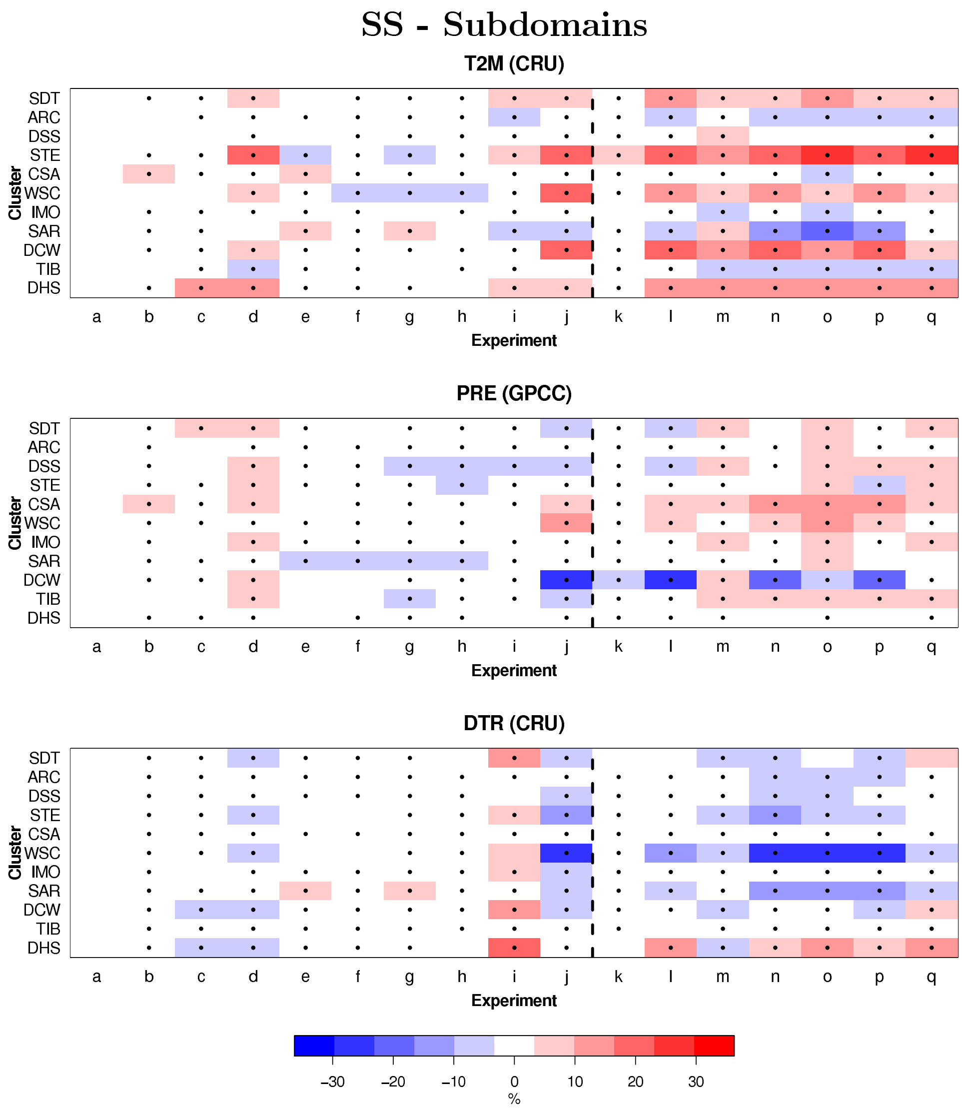

Figure 8: Skill Score (SS) derived from the MAE calculated for each domain sub-region over the monthly climatological values of the seasonal cycle of different COSMO-CLM simulations and observational datasets. From top to bottom, each panel represents the results obtained for near surface temperature (T2M) using the CRU, for precipitation (PRE) using the GPCC and for diurnal temperature range (DTR) using the CRU as observational datasets. Positive (negative) values indicate better (worse) performance of the considered simulations compared to the reference one. The points in correspondence of different clusters and experiments indicate that the change is the SS are the same in sign, when considering all the observational datasets for each variable.

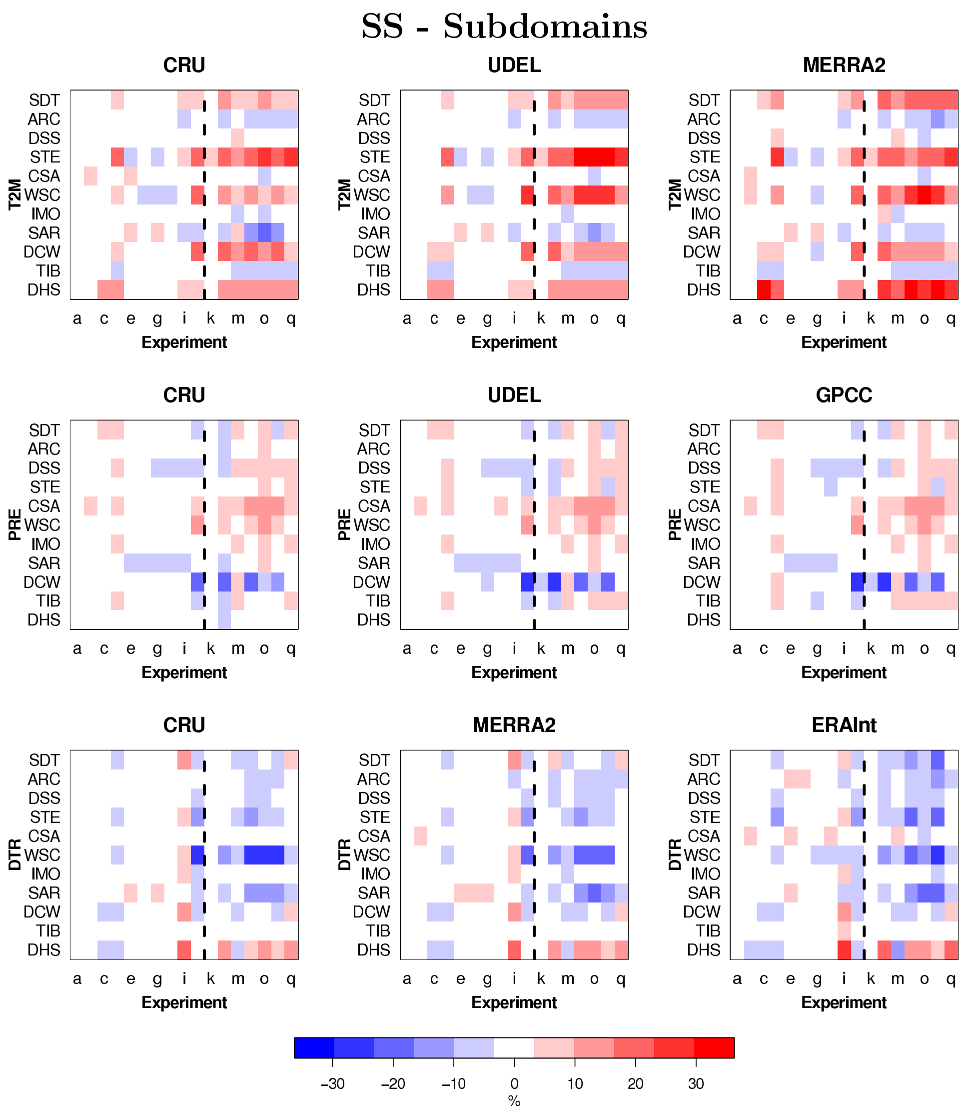

Figure 9: Skill Score (SS) derived from the MAE calculated for each domain sub-region over the monthly climatological values of the seasonal cycle of different COSMO-CLM simulations and observational datasets. From top to bottom, each panel represents the results obtained for near surface temperature (T2M) , for precipitation (PRE) and for diurnal temperature range (DTR). In each case the results obtained for different observational datasets are shown. Positive (negative) values indicate better (worse) performance of the considered simulations compared to the reference one.

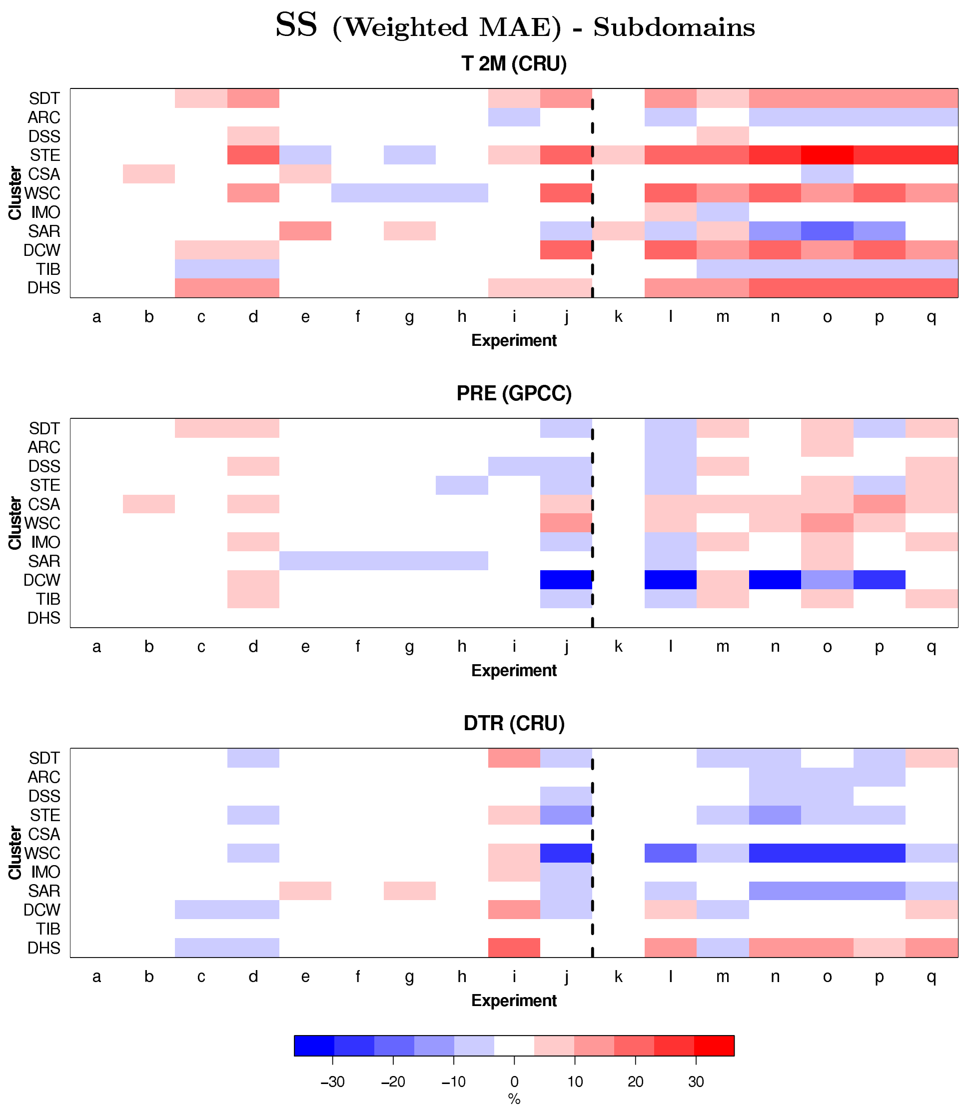

Figure 10: Skill Score (SS) derived from the weighted MAE (MAE_w, Eq. 1) calculated for each domain sub-region over the monthly climatological values of the seasonal cycle of different COSMO-CLM simulations and observational datasets. From top to bottom, each panel represents the results obtained for near surface temperature (T2M) using the CRU, for precipitation (PRE) using the GPCC and for diurnal temperature range (DTR) using the CRU as observational datasets. Positive (negative) values indicate better (worse) performance of the considered simulations compared to the reference one.

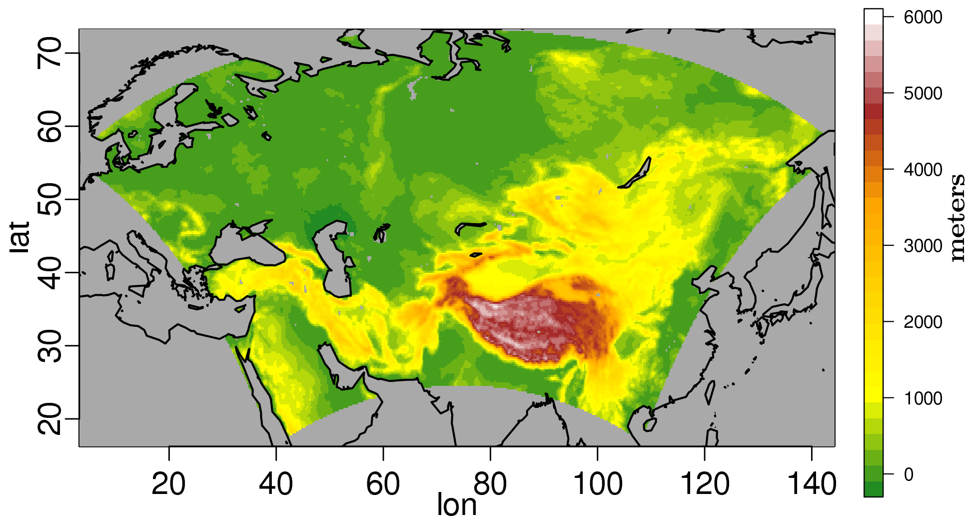

Figure 11: Orography map of the Central Asia simulation domain on a regular grid with a spatial resolution of 0.25 km. Masked in gray are the ocean and the external area of the domain.

With kind regards on behalf of the all authors,

Emmanuele Russo