Dear Editor and Reviewers,

Many thanks for your valuable comments and suggestions, which were very useful for improving presentation of results and paper readability. In the following, we will answer all the comments raised by the reviewers in detail. The reviewers' comments are in bold, citations in Italic and our answers are in regular font.

1 Reviewer 1:

1.1 General comments:

Overall, the authors present some interesting results that merit to be published. They explored a traditional method of downscaling and contrasted it with a more novel machine-learning approach, which is interesting and a hot topic in the field of climate modeling.

We thank you for your positive and encouraging comments. We appreciate your effort to review our manuscript with such a detail and believe that your critics improved the results presented a lot. Especially, your many helpful comments in the PDF helped us to modify the text easily at the right position.

The experiments and results only need minor modifications: the method needs more transparency and explanation of the authors' choices. Their experiments are not reproducible from their text but could be with some additional information.

We agree on this point. We have moved the code and the description from GitHub to Zenodo, where it received a DOI and provided the input datasets as well as trained models there. Regarding the methodology, we have added two new schematic figures explaining the general work done in this manuscript and the architecture of the CNN used in the deep learning approach in details (new figures 1 and 2).

The text needs major rewriting and restructuring as the manuscript sent for review really lacks quality. The manuscript still contains several typos, grammar errors, inconsistencies, missing references, and things in the wrong place. This manuscript did not read as "review ready," and it could have strongly benefitted from additional proofreading before sending it to reviewers. Furthermore, a sentence in the introduction was clearly copied from another website without credit. This is unacceptable and should not have happened, as it comes across as sloppy and unethical.

Thanks a lot for finding the similarity of the text which was ignored completely by the Copernicus software. We removed it from the new version of the manuscript and tried to introduce CORDEX in a way that the similarity index is 0 to the one of their website and any published paper.

Therefore, I advise the authors to thoroughly rewrite and clean their manuscript before sending it back for review.

We have modified the manuscript and rewrote it according to your comments.

Specific comments:

Answers to reviewing questions from Copernicus:

1- Does the paper address relevant scientific modeling questions within the scope of GMD? 2- Does the paper present a model, advances in modeling science, or a modeling protocol that is suitable for addressing relevant scientific questions within the scope of EGU? 3- Does the paper present novel concepts, ideas, tools, or data? Does the paper represent a sufficiently substantial advance in modeling science?

A1+2+3 together: The paper presents the added value of dynamical downscaling precipitation from a GCM over Central Asia and compares it to emulating the RCM with a machine learning framework. Although statistical downscaling with machine learning has been done with precipitation (and other climate variables) before, it seems novel over Central Asia, and the authors make some interesting comparisons. I think the geoscience community can benefit from learning more about the benefits of these machine-learning models and how they compare to traditional techniques.

Thanks for the positive comments.

4- Are the methods and assumptions valid and clearly outlined? 5- Are the results sufficient to support the interpretations and conclusions? 6- Is the description sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? In the case of model description papers, it should, in theory, be possible for an independent scientist to construct a model that, while not necessarily numerically identical, will produce scientifically equivalent results. Model development papers should be similarly reproducible. For MIP and benchmarking papers, it should be possible for the protocol to be precisely reproduced for an independent model. Descriptions of numerical advances should be precisely reproducible.

A4+5+6 together: At the current state of the manuscript no. The methods are not clearly outlined, the results need clarification and as it is, the experiments could not be reproduced by fellow scientists. But if the manuscripts are rewritten with the proposed corrections and feedback, I believe that it could be.

We answered all specific detailed comments raised by you in the new version of the manuscript. We added 2 new figures only explaining the methodology. We re-plotted all the figures according to your comments. We checked the Zenodo Repository if one could reproduce the same results from scratch.

7- Do the authors properly credit related work and clearly indicate their new/original contribution? A: No, at the current state of the manuscript, there is a copying problem from a website in the introduction that needs crediting or paraphrasing.

As mentioned previously, we solved this.

8- Does the title clearly reflect the contents of the paper? Yes.

9- Does the abstract provide a concise and complete summary? Yes.

10- Is the overall presentation well-structured and clear? A: No. The majority of sections need restructuring and rewriting to make it clear.

We have restructured and rewrote a large portion of the manuscript.

11- Is the language fluent and precise? A: No. There are typos and grammar errors that need to be corrected.

Done.

12- Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? A: No, equation 1 needs verification. There is also a lack of consistency for acronyms across the manuscript.

Done.

13- Are the number and quality of references appropriate? Yes.

Is the amount and quality of supplementary material appropriate? For model description papers, authors are strongly encouraged to submit supplementary material containing the model code and a user manual. For development, technical, and benchmarking papers, the submission of code to perform calculations described in the text is strongly encouraged. A: Yes.

Section by section:

Abstract: Overall, you wrote a good abstract. It generally reads well and gives a good overview, but it needs some clarifications and minor changes.

Done.

Introdution:

- **Acronyms: Your text has many acronyms; this weighs down the text and makes it hard to read. Some are used only once or twice, so I suggest you remove those and keep only essential acronyms. There is also the problem of using acronyms before they are defined (usually, only later in the text); that should be corrected.**

  Done.

- **Structure: The introduction feels scattered and all over the place. Information is repeated, and some paragraphs talk about different subjects. Some paragraphs also contain sentences that feel out of place and should be in another place (highlighted in the annotated pdf). The introduction needs to be rewritten and restructured so that similar information is not repeated and appears in the same place in the text.**

  According to your annotated PDF, we have changed the introduction a lot! We removed unnecessary information and focused on two main scientific questions presented by bullet points in the introduction.

- **Quality: the sentence on lines 55-57 was clearly copied from the Cordex website (see annotated pdf). The two sentences (from your text and the Cordex website) have a Jaccard similarity index of 0.69, but it was not picked up by Copernicus' similarity report. This is unacceptable and very sloppy. You should modify this immediately so that you paraphrase it in a way that is not just a copy.**

  We modified this part and checked the similarity index with an online software. We have reached 0% similarity in the new version of the manuscript.

Data and methods:

- **Overall, this section needs restructuring and rewriting. There are inconsistencies, and quite a few pieces of information need to be included or clarified. You might understand your setup very**

well, but it's hard to follow exactly what you've done for a new reader, creating a transparency problem.

We absolutely agree on this. We hope with the new schematics and explanations, especially the added parts on the CNN, we reached a quit good level of the transparency.

- **The CNN section particularly needs considerable rewriting and more information about your choice of the framework (i.e., what's the perfect and imperfect framework and why did you choose the imperfect over perfect one), the selection of training/validation/testing data (and why), and the different models (HCL, No-CL, SCL) you created (see annotated pdf for more details). Things from the appendix need to be in the CNN section, and your different CNN setups (NoCL, etc.) must be clearly defined there. I think it's also missing a figure for the architecture, and readers would benefit from seeing it in your manuscript instead of having to look it up in another paper.**

Thanks to your annotated PDF, we implemented the missing information and clarified this part in the new version of the manuscript.

- **You should also consider what your target audience will be for this paper. Because you aim for a geoscience journal, its readers might not necessarily be familiar with many machine learning terms and might need more background information. I think your paper might benefit from more explanations like: why did you choose to use a CNN instead of another architecture? What are its advantages? Why did you choose to train it in this way?**

we added the following paragraph to the introduction :

"CNN can recognize and encode spatial hierarchies in data (Zhu et al., 2017), making them exceptionally suitable for geospatial data, which is fundamental in climate modelling. Unlike traditional statistical methods that often require manual selection and careful engineering of features, CNN automatically learns the most predictive features directly from the data (Reichstein et al., 2019). CNN can model complex non-linear relationships between input data and outputs, often present in climate data due to intricate interactions in weather systems. CNN is generally more straightforward and efficient for tasks that aim to predict or classify based on patterns distributed across the spatial domain, such as temperature or precipitation patterns in climate models (Racah et al., 2017). CNN is adept at maintaining spatial coherence in the output, which is critical in downscaling where preserving the geographical patterns of climate variables (like precipitation) is crucial (Kurth et al., 2018)."

and we have added some details on our choice of CNN in the new version of the manuscript.

Results:

- **Overall, your results are interesting, and you conducted some good experiments for your RCM and CNN emulator. For example, I think it's great that you also tested your CNN emulator on a new GCM. However, the way you present your results needs more transparency. Some figures would benefit from additional information to be clearer (see annotated pdf). Also, your interpretations rely greatly on seeing "by eye" how different the patterns are and how well the emulator reproduces the RCM. Some of them also come across as vague and unconvincing. Adding a quantitative value to this (for example, spatial means and std of AV or MAE) could add weight to your arguments.**

  We re-plotted all the figures to meet your critics. We also added one additional figures showing simple biases along with the MAE. The boxplots are thought to present the quantitative values, which resent medians and standard deviations.

- **Quality: This section also needs rewriting. Some things need to be put in the right place, lack transparency, a reference is missing (? from Latex), and there are inconsistencies in notations.**

  Thanks to your detailed critics, we have done many corrections on these points.

- **Question to the authors: From what I understood, you ran your CNN over a dataset that combines the historical period and different SSPs. In the hypothesis that you randomly distributed your training and testing data, have you evaluated whether the CNN tests differently over the three SSPs? One thing to do, for example, that could have been done to test generalization, for example, could have been to train the CNN on historical + two SSPs and then test it on an unseen SSP.**

  Yes, we shuffled the data. Our motivation was to train the CNN on as many situations as possible. We wished to have different scenarios and find the mapping in many different forcings. However, in the generalization part, meaning, applying on a new GCM, we only did it on one SSP (SSP370). One reason, why we did not

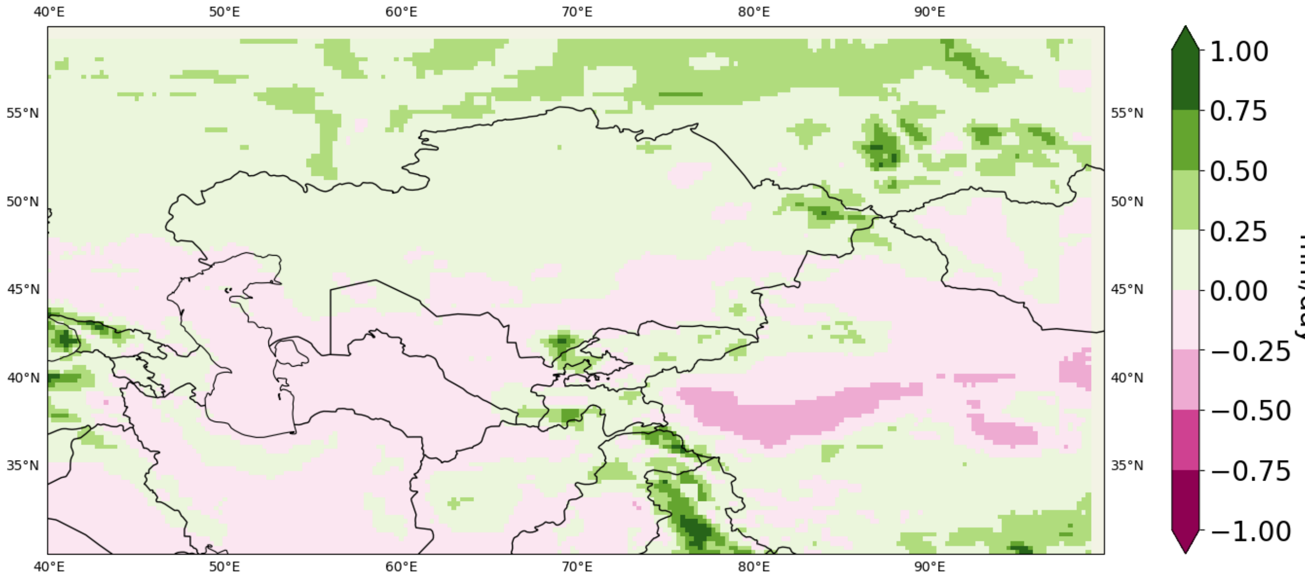

Figure 1: Added Value [mm/day] for the CNN with SSP370 scenario excluded from training.

test the setting you proposed, was that if we leave one SSP aside, then we will reduce the size of training set by almost 30% and we already are at the limits of sample size for deep learning. One possible solution for that might be to redo all the experiences with 3-hourly model output, which is available in our RCM output. Then we might have more data available for training.

For the sake of curiosity, we have followed your generalization suggestion and excluded the SSP370 from the training and the calculated the trained model's MAE on the simulation under SSP370. The AV pattern (Figure 1) looks very similar to the one from the previous model. However, negative AV values are a bit larger than the other CNN.

We inserted this result in the new version of the manuscript as a kind of generalization test on different forcing. This additional test is actually a very interesting approach of generalization. Thanks again for the comment.

2 Reviewer 2:

2.1 Major comments

The study presents results from two downscaling methods for GCM precip itation. One is dynamical downscaling with the CCLM RCM and the second is a statistical emulator for CCLM based on CNNs. The downscaling is applied to simulations with the MPI−ESM1−2_HR GCM. Both approaches are in principle useful and the results are informative.

Thanks a lot for the positive feedback. We have to mention that we also conducted a simulation where we downscaled EC−Earth3−Veg GCM using CCLM and the trained CNN (see Sec. 4.1 Applying CNN to a different GCM and `https://zenodo.org/records/10417111`). We will add more details of the simulations in a new version of the manuscript.

However, the method evaluation is very limited as it is only based on MAE, while many other evaluation measures could have been used. I suggest to add at least an analysis of the bias in the mean and include a justification for the specific choice of evaluation measures. Moreover, the analysis of MAE of GCM simulations and GCM-driven RCM simulations relative to observations is fundamentally wrong, because the random internal variability in the simulations and in the observations is not synchronised.

We are aware that he simulations are in the so-called "free" mode and do not include any kind of data assimilation and do not "see" the observations. However, we conduct averaging of 30 years on each day, i.e. we have 30 first of January for example and the resulted pattern is not only a random pattern of a single day. However, as suggested by you, we added the new figure 5 in the new version of the manuscript, showing biases of the climatological values. With such figures we again confirm the added values of the RCM downscaling conducted in our study for mean states. The overall added value pattern is similar in both cases, i.e. using MAE or climatological bias. This might be embedded in the 30-years averaging in the MAE calculation, which cancels out small random variablities, which are induced by the dynamical downscaling.

The choice of using MAE for the comparisons are motivated by the fact that a mean of absolute errors over 30 years already cancels out the majority of randomness within the dataset. And we are ultimately interested in the reduction of the mean error via downscaling. Other reason for selection of MAE lays in the distribution of

Precipitation. According to the study of Hodson (2022), MAE is more suitable for precipitation than other metrics.

For the rest of the manuscript, we will show the MAE, because we are comparing the emulated pattern against the RCM-pattern and there is no comparison to observations.

The manuscript is also not well written, and important conceptual basics and technical details are not clearly explained. It needs substantial rewriting before it is suited for publication.

We have implemented your critics in the new version of the manuscript and modified the text a lot to meet your expectations.

2.2 Specific comments

The introduction contains many good points, but there is some repetition and it should be better structured. The purpose of the paper should be made clear in the abstract and at the beginning of the introduction. What are the research questions?

We modified the introduction accordingly with a focus on the research questions listed as bullet points in the introduction.

- Research Question 1: How effectively can CMIP6 models be downscaled for the CORDEX Central Asia region to enhance precipitation simulations? (Kendon et al., 2014; Demory et al., 2020; Hess et al., 2022)
- Research Question 2: Can convolutional neural networks (CNNs) effectively downscale GCM outputs, and how do they perform when applied to GCMs they were not initially trained on? (Sun and Lan, 2021; Rasp and Lerch, 2018)

Is this mainly a methodological study or is the main purpose to provide high-resolution scenarios to inform impact and adaptation studies? The first part of the introduction suggests that the CCLM predictions are the main point, and only after line 110 it is said that a ML emulator will be tested, and that three research topics are addressed. The second research topic (line 118) is unclear. What is meant with the 'dynamical downscaling signal for heavy precipitation'? Signal of what? The third topic is training a CNN-based emulator for CCLM, but why is evaluation not mentioned? It would also be good to discuss whether there are already published findings on the added value of RCMs over Central Asia, for instance from CORDEX.

We have modified the introduction a lot to make those points clear. We added a paragraph listing the few downscaling efforts done in Central Asian domain.

Line 11: 'we downscale CCLM' is not correctly phrased

We have changed it to "Additionally, we employ the CCLM to refine future climate projections."

Lines 16-18: The setup of the CNN training and evaluation, and the applications are not clearly explained in the abstract. Of course, the CNN emulator is model-specific, as it is designed to emulate a specific RCM.

We reformulated this part in the new version of the manuscript.

Line 25: Maraun and Widmann (2018) 'Statistical downscaling and bias correction for climate research', Cambridge University Press, is a standard reference for statistical downscaling and should also be cited.

That is true. We added this source in the new version.

Lines 101-102: If ML is used for postprocessing large-scale data separately for each time step, as is the case for the emulator presented, there is no iterative use of the output. It is therefore unclear how this comment relates to the study.

We have removed this sentence and explained the conservation process of the DL architecture we have used in this study. For example the following part is added to the new introduction:

"CNNs are adept at maintaining spatial coherence in the output, which is critical in downscaling where preserving the geographical patterns of climate variables (like precipitation) is crucial (Kurth et al., 2018). Constrained CNNs integrate physical constraints or laws directly into the training process. The constraining is done by changing the loss function or the network's architecture to enforce compliance with physical laws (i.e., conservation of mass, energy, or momentum). Unconstrained CNNs operate without explicitly incorporating physical laws or constraints into the network's architecture or loss functions. They focus solely on learning from the input data to the output predictions based on the data-driven patterns they detect. This study explores unconstrained and constrained CNN approaches to understand their effectiveness in downscaling and how they perform when applied to GCMs on which they were not initially trained."

Lines 153-159: It is not clear what is meant with 'high/low challenges to adaptation'.

we added the following paragraph to the new introduction: "Challenges to adaptation refer to the degree of difficulty that societies might face in adjusting to the environmental, economic, and social impacts of climate change.

Specifically, this term refers to a society's fundamental susceptibility and the accessibility and efficacy of technologies and approaches designed to lessen the impacts of climate change. The adaptation challenges are minimal in the SSP126 scenario, which envisions a sustainable future. This implies that, under this scenario, global cooperation and sustainable practices lead to advancements in technology and governance that significantly reduce vulnerability to climate change impacts. Additionally, societal structures are resilient, and resources are managed to minimise environmental stresses and maximise human well-being."

Section 2.1.1: There should be a reference to Fig. 1a to specify the CCLM domain, and in the figure caption 'study region' should be replaced with 'CCLM simulation domain'.

We added this into the new version.

Section 2.1.2: The setup for the CNN training is not fully clear. What are the simulation periods for the scenario runs? Are the input and output variables daily precipitation?

We implemented this data in the new version of the manuscript.

Line 161-162: If input and output are precipitation, what is the meaning of energy and mass conservation in this context? Is 'energy' and 'mass' in this context the same or are these different quantities? It turns out later that the details are given in appendix A and that for the hard constraint the meaning of 'mass and energy conservation' is that the precipitation over a GCM gridcell is conserved in the high-resolution precipitation. However, the main part needs to be self-contained and written such that it is not confusing. Therefore, a short explanation of the main aspects of the constraints and a reference to the appendix should be given here.

We agree with this point and therefore moved this part to the main part of the manuscript. We explain each layer in the main part of the paper with references to the new figure 2. Description of the constraint layers are now presented in section 2.1.3.

Appendix A is very unsystematic and unclear. Specific problems are listed in the next five points. Line 417: The simplest way to ensure mass conservation would be to scale all small-scale values within a given large-scale gridcell with the ratio of the large-scale value and the sum of the small-scale values. Why is this not done and what is the reason for the specific choice using the exponential dependency of the scaling factors on the small-scale values?

Thanks a lot for this point, this might need further explanation. The way of scaling that you are describing is also one possibility described in Harder et al., 2023. The work showed though that using the softmax constraints layer gives better results. The exponential both ensures positive predictions and leads to more variance between subpixels in the super-resolved prediction. The multiplicative rescaling you are describing e.g. struggles when the sum of the small-scale values gets close to zero.

Line 418: Why is the MAE loss function mentioned in the context of the hard constraint, which in the way it is formulated does not depend on the loss function for the CNN? It is said already in line 192 that MAE is the loss function for the CNN.

That's correct, the hard constraints layers dont depend on the loss function. There is a formulation of the soft constraining, which includes an addional term in the loss function. We have removed that part in the new version of the manuscript.

Lines 419-424: Is this the loss function for the whole CNN, or is it only relevant for the constraint layer? If it is the former, this contradicts the statement in line 192. If it is the latter, the use of two loss functions needs to be explained. How are the y_i calculated from the values in the previous layer? Why is there an explicit version for calculating y_i for the hard constraint (eqn. A2) but not for the soft constraint?

The loss function is always for the whole CNN. If there are hard constraints are included there is no change in the loss function at all. Next to hard constraints there also exist soft constraints, which is a way of enforcing constraints through an additional term in the loss function. Soft constraints and hard constraints are fundamentally different as hard constraints build on reajusting the outputs during training and inference at the end of the neural network through a non-trainable layer. This is an explicit function that is applied. Soft constraints are not enforced explicitly but through the penalizing term mentioned above.

Lines 425-426: It is not clear why a loss function is mentioned if there is no constraint layer. It is already said in line 192 that MAE is the loss function used for the whole CNN.

The loss function is mentioned because it is used for the soft constraints and changed for this.

Lines 428-429: This sentence says that MAE is an evaluation criterion for the different settings. This is confusing. Are the 'loss function' and the 'evaluation criterion' used differently? If so, it needs to be explained how, or the statement on the evaluation criterion needs to be moved to the 'metrics' section.

The MAE is both used as a loss function and an evaluation metric. A loss function is used during the training to optimize the neural networks parameters, while an evaluation metric is calculated on the validation or test data set to evaluate the model on an independent dataset. Those are two different use cases, but both can use an MAE. This information is added into the new version for clarifying.

Lines 162-171: Although GCM errors affect the output, the emulator is a statistical model, which should be not very sensitive to the states used for fitting, otherwise there is the usual problem of stability of statistical relationships in statistical downscaling. The phrasing 'introduce biases in the downscaling process' is misleading, as the 'downscaling process' is the CCLM model or the CNN emulator, not the output. The discussion in this part is conceptually unclear as in conflates models and outputs. The fact that biases and errors of the CCLM and CNN output are partly caused by propagation of GCM biases and errors needs to be taken into account in the evaluation.

We changed the whole paragraph into a new one :

"In the context of deep learning for climate modelling, the 'perfect model' approach involves starting with high-resolution data, which is considered accurate or nearly perfect, and intentionally degrading it to a lower resolution. The aim is to simulate a scenario where the 'truth' (the original high-resolution data) is known, and then to recover this high-resolution from the artificially degraded data using deep learning techniques. This approach is a crucial part of training, as it teaches the model the desired mapping from low to high resolution, enabling the model to effectively learn how to upscale or enhance resolution while minimizing the loss of critical information. It's a controlled experiment that helps refine the model's capabilities.

The "imperfect model" approach, on the other hand, acknowledges that both the low-resolution (GCM output) and the high-resolution (RCM output) datasets have their inherent errors and limitations. In this scenario, we do not have a single source of truth but rather two separate sets of data:

- Low-resolution data: This may capture global or large-scale phenomena but miss regional details (Xu et al., 2021; Chokkavarapu and Mandla, 2019).

- High-resolution data: This provides detailed regional information but may still have errors or not perfectly reflect reality due to limitations in data collection, model configuration, or computational constraints (Muttaqien et al., 2021).

In this setup, the challenge for deep learning is to learn a mapping between these two independently imperfect data sets. With using the CNN we try to train a model that can predict high-resolution details from low-resolution inputs as accurately as possible despite the absence of a perfect ground truth. This involves understanding and modeling the uncertainties and biases inherent in both datasets."

Lines 189-190: The statement that the unconstrained CNN works best is based on the evaluation, and the performance ranking depends in principle on the evaluation measures. This is the method section and the reader does not know yet what the evaluation measures are. The statement on the best architecture should be moved to the result section and it should be made clear that the ranking of methods can depend on the evaluation measures.

We agree and have moved this part to the results and explain it in a right order now.

Lines 196-206: How different are the CHIRPS, APHODITE and CPCC data on the coarser grids for which they are all available?

Based on the point raised by reviewer 1, we have removed this lines from the new version of the manuscript.

Lines 210-215: It would be good to also do the evaluation on the coarse grid, or at least comment why this is not done. See also Volosciuk et al. (HESS 2017) 'A combined statistical bias correction and stochastic downscaling method for precipitation' for a discussion on the distinction between biases and differences in statistical properties on different spatial scales.

We did not initially include an evaluation on the coarse grid primarily due to our specific focus on the impacts at a higher spatial resolution, which is more relevant for local climate adaptation strategies, like the work of Russo et al. (2020). Nevertheless, to enhance the robustness of our findings and provide a comprehensive analysis, we will include a discussion on the implications of not evaluating at coarser resolutions in the revised manuscript. This discussion will reference the findings of Volosciuk et al. (2017) to provide a theoretical framework for understanding the scale-dependent nature of biases and differences in statistical properties.

We added the following paragraph to the discussions : "Our study evaluated the downscaling skill primarily using higher resolution observations, which are critical for capturing localized climate phenomena relevant to regional adaptation strategies. However, as Volosciuk et al. (2017) noted, examining downscaling outputs at coarser resolutions can be equally informative. Their work emphasizes that downscaling methods can introduce or fail to correct biases that differ significantly across spatial scales. By evaluating on a coarser grid, it is possible to distinguish between the inherent biases of the model and those introduced by the downscaling process. This distinction is crucial for understanding the limitations and strengths of downscaling methods in representing climatic variables across different scales."

Lines 216-223: The evaluation is very limited because it only addresses temporal variability, and

only with one specific measure. Other measures for the agreement of simulated and observed temporal variability could also be used such as correlations of the timeseries or Brier Score for threshold exceedances. Differences in distributions, including the bias in the mean, potentially also in quantiles should also be analysed. The various aspects of evaluation are discussed for instance in Maraun et al. (Earth's Future 2015) 'VALUE: A framework to validate downscaling approaches for climate change studies'.

Indeed we show correlations as numbers in the paranthesis within new Figure 9, new Figure 10 c and d, whenever we do not shuffle the datasets. We also added the simple bias maps in new figure 5. The current model set-up has already been evaluated in a previous study i.e. Russo et al. (2020). Here, we apply that setting for creating the CMIP6-based CORDEX-CA.

Moreover, the terminology is not correct. A bias is a systematic difference between a statistical variable calculated from two datasets. It often refers to variables that characterise distributions (such as the mean, variance, or quantiles), but can also be used for variables that characterise temporal variability (such as autocorrelation or spectra) or spatial variability (such as correlation lengths, for instance Widmann et al. (IJC 2019) 'Validation of spatial variability in downscaling results from the VALUE perfect predictor experiment'. It is not common practice to use the term bias to characterise the agreement of individual time steps, and therefore MAE should not be called bias.

This issue was raised also by reviewer 1 and we agree with it. We now show both MAE and the simple bias maps. Please see answer to reviewer 1.

Section 3.1.1: If I understand correctly the MAE is calculated based on pairs of daily simulated and observed values. If so, this approach is fundamentally wrong, because the precipitation series are realisations of random internal variability, which are different in the observations and in the GCM simulations or GCM-driven RCM simulations. This is different for reanalyses and reanalysis-driven RCM simulations because of the data assimilation in reanalyses. MAE is based on pairs of values for a given time and a measure for how similar the timeseries are. It makes no sense to calculate MAE for non-synchronised timeseries, because there is no justification for the paring of values. In this situation the MAE is only affected by the difference in variance and provides no meaningful measure of agreement of the specific temporal behaviour. This section and Fig. 3 should therefore be deleted.

We completely agree on this point and removed the subsection and its figures.

Lines 259-261: The GCM biases that affect the climate change signal in RCM simulations are not the MAE for short-term variability but systematic biases for instance in the large-scale mean circulation for the present climate, which may be linked to unrealistic large-scale climate change. The links between biases and climate change signals are complex, and should be discussed more carefully.

As mentioned above, this section is removed from the new version of the manuscript.

Line 322: Project reference is missing

We corrected that.

Lines 351-353: This argument is missing the main point. CCLM is driven at the lateral boundaries by the GCM values for the state variables of CCLM (temperature, pressure, wind speed etc.). Precipitation is not used for driving the RCM. The CNN input is the GCM precipitation, which has different biases in the two GCM, and therefore the mapping from the MPI-GCM-precipitation to the CCLM precipitation cannot be successfully transferred to EC-Earth.

We added this point in the new version.

Discussion and conclusions: This section needs to be rewritten after the issues listed above have been addressed.

We have rewritten those sections.

Line 516: Reference for the Harder et al. 2022 preprint should be updated to the peer-reviewed version Harder et al. 2023.

Done.

On behalf of all authors,

Bijan Fallah

References

Chokkavarapu, N. and Mandla, V. R.: Comparative study of GCMs, RCMs, downscaling and hydrological models: a review toward future climate change impact estimation, SN Applied Sciences, 1, 1698, 2019.

Demory, M.-E., Berthou, S., Fernández, J., Sørland, S. L., Brogli, R., Roberts, M. J., Beyerle, U., Seddon, J., Haarsma, R., Schär, C., et al.: European daily precipitation according to EURO-CORDEX regional climate models

(RCMs) and high-resolution global climate models (GCMs) from the High-Resolution Model Intercomparison Project (HighResMIP), Geoscientific Model Development, 13, 5485–5506, 2020.

Hess, P., Drüke, M., Petri, S., Strnad, F. M., and Boers, N.: Physically constrained generative adversarial networks for improving precipitation fields from Earth system models, Nature Machine Intelligence, 4, 828–839, 2022.

Hodson, T. O.: Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not, Geoscientific Model Development, 15, 5481–5487, 2022.

Kendon, E., Roberts, N., Fowler, H., Roberts, M., Chan, S., and Senior, C.: Heavier summer downpours with climate change revealed by weather forecast resolution model. Nat. Climate Change, 4, 570–576, 2014.

Kurth, T., Treichler, S., Romero, J., Mudigonda, M., Luehr, N., Phillips, E., Mahesh, A., Matheson, M., Deslippe, J., Fatica, M., et al.: Exascale deep learning for climate analytics, in: SC18: International conference for high performance computing, networking, storage and analysis, pp. 649–660, IEEE, 2018.

Muttaqien, F. H., Rahadianti, L., and Latifah, A. L.: Downscaling for Climate Data in Indonesia Using Image-to-Image Translation Approach, in: 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 1–8, IEEE, 2021.

Racah, E., Beckham, C., Maharaj, T., Ebrahimi Kahou, S., Prabhat, M., and Pal, C.: Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events, Advances in neural information processing systems, 30, 2017.

Rasp, S. and Lerch, S.: Neural networks for postprocessing ensemble weather forecasts, Monthly Weather Review, 146, 3885–3900, 2018.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, f.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, 2019.

Russo, E., Sørland, S. L., Kirchner, I., Schaap, M., Raible, C. C., and Cubasch, U.: Exploring the parameter space of the COSMO-CLM v5. 0 regional climate model for the Central Asia CORDEX domain, Geoscientific Model Development, 13, 5779–5797, 2020.

Sun, L. and Lan, Y.: Statistical downscaling of daily temperature and precipitation over China using deep learning neural models: Localization and comparison with other methods, International Journal of Climatology, 41, 1128–1147, 2021.

Volosciuk, C., Maraun, D., Vrac, M., and Widmann, M.: A combined statistical bias correction and stochastic downscaling method for precipitation, Hydrology and Earth System Sciences, 21, 1693–1719, 2017.

Xu, Z., Han, Y., Tam, C.-Y., Yang, Z.-L., and Fu, C.: Bias-corrected CMIP6 global dataset for dynamical downscaling of the historical and future climate (1979–2100), Scientific Data, 8, 293, 2021.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources, IEEE geoscience and remote sensing magazine, 5, 8–36, 2017.