Dear Editor,

Thank you for your valuable feedback. We have carefully considered both of the major comments raised in the revision process and addressed all of the reviewers' questions in detail in the updated version of the manuscript.

0.1 General comments:

A major revision is needed to improve the scientific quality and presentation quality of this manuscript. As pointed out by one of the reviewers, the conceptual clarity of the evaluation approach and the linkages between different parts of the study need to be addressed. MAE reductions can be caused by different reasons, such as by reductions in the climatological mean bias, by reductions in the bias of the variance, by improvements of correlations between simulated and observed time series, or by a combination of these factors. The authors should consider including multiple metrics to evaluate different aspects of the model and improving the coherency of the manuscript.

To support the MAE reductions reported in our CMIP6-CORDEX-CA simulations, we have incorporated additional analyses. Specifically, we now evaluate the model's performance using multiple metrics, including MAE, RMSE, climatological bias, correlation, comparison of probability density functions (PDFs) of maximum daily precipitation (provided in the Appendix), and the Relative Probability Difference Index. These additions provide a more comprehensive assessment of the error reduction mechanisms.

Furthermore, we conducted a comparative analysis of the two GCM simulations used in our study to explore their similarities.

We discussed why the unconstrained CNN might have performed better.

In the revised manuscript, we have also provided an extended discussion on the caveats, strengths, and future outlook of the work.

On behalf of all the authors, Bijan Fallah

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 Minor comments

Good job for the improvement in the writing quality of your manuscript; it's now a lot better! I spotted some last mistakes and missing words that you'll find in my annotated manuscript with comments (see annotated pdf). The other comments in the annotated pdf apply to results and discussion, which need some more work to present your results fairly.

Thank you very much for your valuable comments. We will first address the comments provided in the PDF file, followed by a response to the remaining feedback:

P1: be careful with using the terms significant as it requires a statistical test of significance. If you didn't do that I would use another word that is a synonym.

Done.

P1: it's not clear here what this value represents and I also cannot find it in the results. Is it a difference between in MAE between the RCM and GCM or something else? This should be specified.

That is indeed very true. Thanks for hinting that. We changed the text to clarify the reduction in MAE and Bias during summer, winter and annual. Also we added some analysis about the frequency of maximum daily values to the Appendix A. That analysis is also linked to the editor's comment. We show that higher resolution datasets show maximum daily values nearer to the CHIRPS dataset and the Global models or reanalysis are not able to do so.

P1: see comment above, and provide a number.

We changed the text accordingly. We added the following information to the abstract : "The number of days with precipitation exceeding 20 mm increases by more than 90 by the end of the century, compared to the historical reference period, under the SSP3-7.0 and SSP5-8.5 scenarios. The annual 99th percentile of total precipitation increases by more than 9 mm/day over mountainous areas of Central Asia by the end of the century, relative to the

1985–2014 reference period, under the SSP3-7.0 and SSP5-8.5 scenarios."

P1: mmh you showed that it actually has trouble generalising to a new GCM so how can it have an added value?

Done! The new text is : "The CNN successfully emulates the GCM-CCLM model chain over large areas of CA, but shows reduced skill when applied to a different GCM-CCLM model chain."

P3: that is implied

We removed this sentence.

P6: change this to "imperfect model" setup. The "this" is not clear to which it refers

We used "In the imperfect model setup,"

P6: the Done!

P6: did you use cross validation?

No, CR is not commonly used in deep learning approaches (like CNNs) because it is computationally expensive. Instead, we rely on a single validation set to track model performance. This is due to the large number of parameters in CNNs and the resources required to train these models over multiple folds (Bengio, 2012).

While cross-validation can help with smaller datasets, there is also a risk of overfitting if the model is highly complex (like CNNs). Since CNNs are powerful enough to memorize smaller datasets, using the same data in multiple folds could lead to overfitting as the model is exposed to repeated patterns. In such cases, cross-validation might not prevent overfitting but could actually increase its likelihood.

In our case, if the dataset is small, the risk of overfitting is more related to the capacity of the CNN rather than the absence of cross-validation. Cross-validation helps small datasets but can overfit if the model complexity is too high compared to the size of the data (Goodfellow, 2016).

To demonstrate that the out training step does not trap in this problem, we show the training and validation loss diagram in Figure 1. As can be seen the validation and training loss reaches a steady state at the end of the training process for NoneCL (other models show similar behavior).

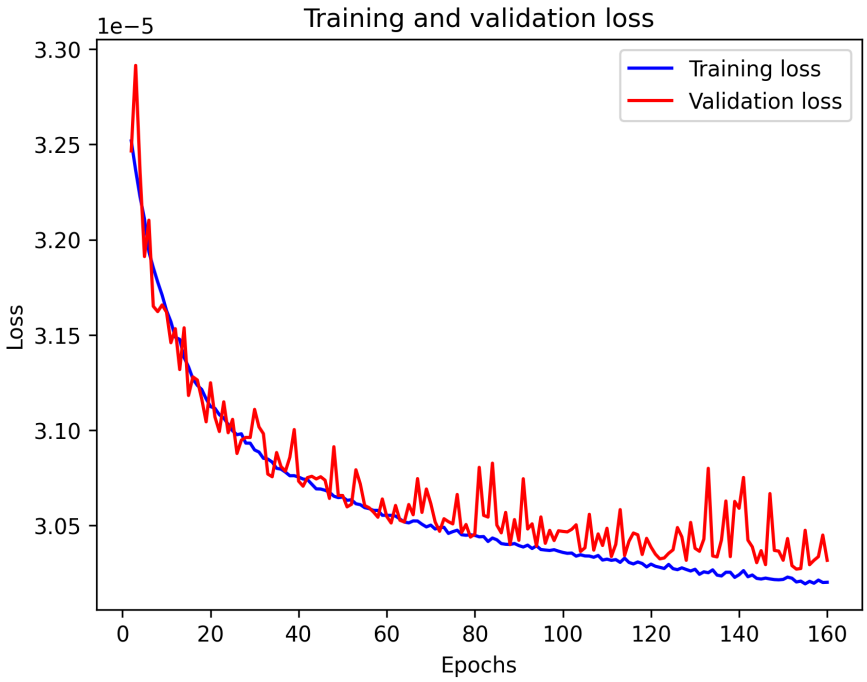


Figure 1: Training and validation loss with respect to epochs.

Overall, this seems like a reasonable training curve for a CNN. The model shows steady improvement on both training and validation data without over-fitting.

P6: independent

Done!

P6: this should follow the order in which they come below

We reordered them.

P7: function

Done!

P7: "Then the mean absolute error (MAE) is applied as a loss function between the final output and the target value."

Corrected!

P7: It's not very clear here if there is a constraint layer like hard constraining or just a special loss.

This approach does not impose a hard constraint through a dedicated layer or architectural component. Instead,

we add a regularization term to the loss function, which allows for soft enforcement of the constraint. The Mean Absolute Error (MAE) loss is extended with an additional constraint violation (CV) loss term.

We clarified that by adding those information in the new version of the manuscript. Thanks for this comment.

P7: The model is trained over

added.

P7 : Can you give a reason why the lr for SCL is so much smaller? There's a factor 100 between them.

It is purely based on the empirical performance (we have tested several *lr* values). A lower learning rate allows the model to fine-tune more carefully in the presence of complex loss terms, leading to better convergence and improved results. This is likely why we observed a better performance with this setup.

Harder et al. (2023) also used different learning rates for their constrained and unconstrained models. While they did not explicitly state a reason for using a much lower learning rate for the soft-constrained loss (SCL), this is a common strategy in deep learning to stabilize training in cases where the additional loss term (such as a regularization or constraint violation term) can lead to more significant gradients or slower convergence.

P7: This is superfluous, especially because evaluation comes later. So you can remove this.

This part was added due to your comment in the previous round (on the usage of MAE). We have removed it from the new version now.

P8: Above you used the term "validation", stay consistent on the terms.

Done!

P8: this sentence is unclear, using CHIRPS directly as what? Are these the 20% of testing stated above in l. 164? Needs rewriting and I think you can remove the "instead of [...] directly

We have modified the paragraph and splitted that into two paragraphs: *"According to Ciarlo et al. (2021), the choice of observational data can significantly influence the perceived added value of an RCM, particularly when detecting extreme events, where poor-quality data might misleadingly suggest improved model performance. They recommend using observational datasets with spatiotemporal resolutions comparable to the model's for enhanced accuracy. In line with this, we use CHIRPS, a high-resolution gridded observational dataset, to validate the CCLM driven by the GCM. CHIRPS offers a resolution of 0.05°, covering latitudes from 50°S to 50°N, and provides independent observations derived from satellite and station data. This contrasts with reanalysis datasets, which rely on climate model simulations (Funk et al., 2015). For the validation of the CNN, however, we allocate 20% of the CCLM*

simulation data as the target for evaluating the CNN emulator's performance rather than directly using CHIRPS. This is because the CNN is designed to emulate the climate output produced by the CCLM, not to match observational data directly. While CHIRPS is used to validate the accuracy of the CCLM output, we validate the CNN by ensuring it accurately reproduces the CCLM's fine-scale climate information, which has already been verified against CHIRPS for its realism.".

P8: Wouldn't F be the same as yi?

We modified that.

P8: as an evaluation metric

Done!

P8: This should be on a new line.

Done!

P8: and this should be F again?

Done!

P8: above this is called O

Done!

P9: can we have a value here? And you are plotting the difference so you don't know if the CCLM's MAE is "high" or "lower", only that it's difference is big.

We are showing the differences of CCLM's MAE and the one from ERAInterim. If we have positive values it means that errors of ERAInterim are larger and 5 mm/day is also a significant difference. And magenta regions show negative values which show better skills for ERAInterim. ERAInterim for example during DJF in figure 4.b has already 2-3 mm/day MAE.

We put numbers to clarify that in more details.

what do you mean with increase and reduction of the MAE? Compared to what? I think you should compare it to the GCM. So in Afgh-Taj, the CCLM is closer to observations than the GCM blabla.

We agree with your comment. We will re-frame the text accordingly to : *"In the mountainous areas of Afghanistan, Kyrgyzstan, and Tajikistan, CCLM is closer to observations than GCM. However, GCM is closer to observations near the domain's southern boundaries throughout the year and in the south and southeast during summer".* In this case GCM is the ERAInterim reanalysis.

Other than that, to demonstrate the skill of our CCLM simulations and also in line with the **Editor's** comment, we conducted the following extra analysis, which might be useful for CORDEX-like model evaluation, especially those focusing on extreme events like in precipitation. Interestingly, the same added value patterns are obtained by the new applied metric, showing our RCM simulation, shows skills similar to a typical CORDEX-CA simulation and ads values over mountains of CA.

We "hope" that doing those analysis will show our simulations might be useful for the community:

We utilize the spatially distributed added value index as introduced in Ciarlo et al. (2021) to evaluate the performance of our RCM simulations. This metric is particularly well-suited as it captures the spatial variability of added value, providing a detailed assessment of how well our model reproduces fine-scale climate features compared to a coarser global climate model (GCM). By focusing on high-resolution regional models, the added value index allows us to quantify improvements in regions with complex topography or localized climate phenomena, which are often missed by coarser models. Moreover, it is an ideal metric for examining extreme events, a key focus in climate impact studies, as it helps highlight where our emulator outperforms the GCM in simulating such events. The spatial nature of this metric also enables us to produce clear visualizations of model performance across various regions, facilitating comparison and communication of results. Overall, the added value index offers a robust and versatile tool for assessing our model's ability to capture regional climate details, which is critical for evaluating the effectiveness of climate model downscaling.

Metrics

We use the formula presented by Ciarlo et al. (2021) to calculate the added value. In the first step, the selected GCM, RCM and observational data are interpolated onto the RCM grid using the distance-weighted average method. Interpolation of the coarser grid to a higher one might create unrealistic values. This issue was discussed in the work of Ciarlo et al. (2021). Then, for each grid point of the domain, we calculate the probability density function (PDF) from GCM, RCM and observation. As in the case of Ciarlo et al. (2021), for a fair comparison, the bin size is fixed to 1mm/day and the maximum value at each grid point for the calculation of the PDF is taken from the maximum of all datasets at that grid point. At each grid point, the absolute differences (D) between the frequency (N) of the model (M) and observation (O) at each bin (ν_t) are divided by the sum of the observation value:

$$D_M = \frac{\sum_{\nu=1}^{\nu_t} |(N_M - N_O)\Delta\nu|}{\sum_{\nu=1}^{\nu_t} (N_O \Delta\nu)}. \tag{1}$$

The added value at each grid point (A_i) is then the difference between the D_{GCM} and D_{RCM}. Positive (negative) values indicate an improvement (degradation) of the downscaling compared to the GCM:

$$A_i = D_{GCM} - D_{RCM}. \tag{2}$$

There shall be a condition added to the added value calculation in the case that $N_{GCM} = 0$ in a bin and N_{RCM} and N_O are nonzero. Under this condition, that bin must not contribute to the D_{RCM} calculation.

Analogous to the A_i calculation, the climate change downscaling signal is presented by :

$$D_{Mf} = \frac{\sum_{\nu=1}^{\nu_t} |(N_{Mf} - N_{Mh})\Delta\nu|}{\sum_{\nu=1}^{\nu_t} (N_{Mh}\Delta\nu)}. \tag{3}$$

where Mf is the future projection, Mh the corresponding historical period and :

$$A_{DS} = D_{GCMf} - D_{RCMf}. \tag{4}$$

Where A_{DS} is the downscaling signal, large positive or negative values indicate a significant climate change downscaling signal. Values near zero indicate a weak downscaling signal.

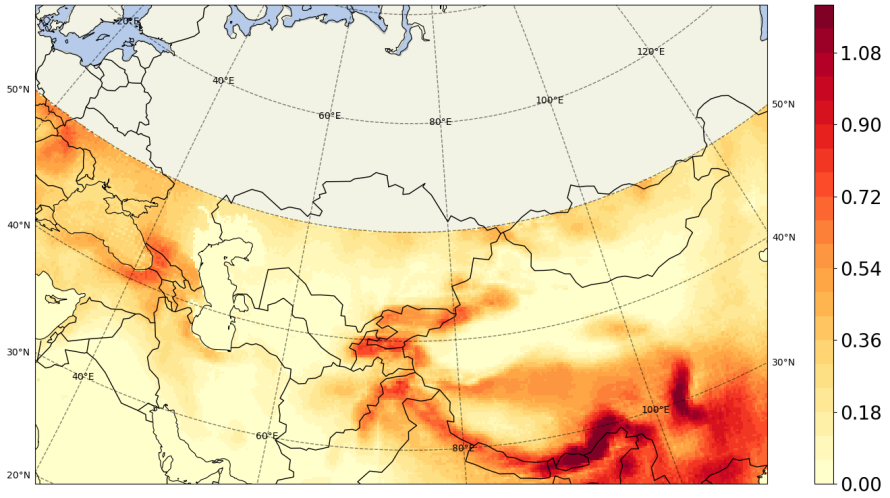
Added value of CCLM driven by ERAInterim

Relative probability differences and added values of ERAInterim reanalysis, CCLM, GERICS-REMO2015 and RMIB-UGent-ALARO-0 driven by ERAInterim (hereafter, ERAInterim, CCLM-ERAInterim, GERICS-REMO2015-ERAInterim and RMIB-UGent-ALARO-0-ERAInterim) are shown in Figure 2. The CHIRPS dataset is used as the observational dataset O to calculate D according to equation 1. For example, D(ERAInterim) indicates the relative probability difference between ERAInterim and CHRIPS. The measure of added value A is more pronounced over areas with complex topography for all three RCMS (Figs.2.e-g). For example, the CCLM and GERICS-REMO2015 models show very negative values north of the Tibetan Plateau, whereas the RMIB-UGent-ALARO-0 model shows positive values over corresponding areas. Over the Southeast of the domain, the CCLM and GERICS-REMO2015 show overall positive added values where the RMIB-UGent-ALARO-0 does not. Considering the whole domain, all three models sensibly reduce the large and local-scale bias of ERAInterim against CHIRPS (Fig.2), especially over complex topographies. Usually, the nested RCMs show similar values of D near their lateral boundaries, with

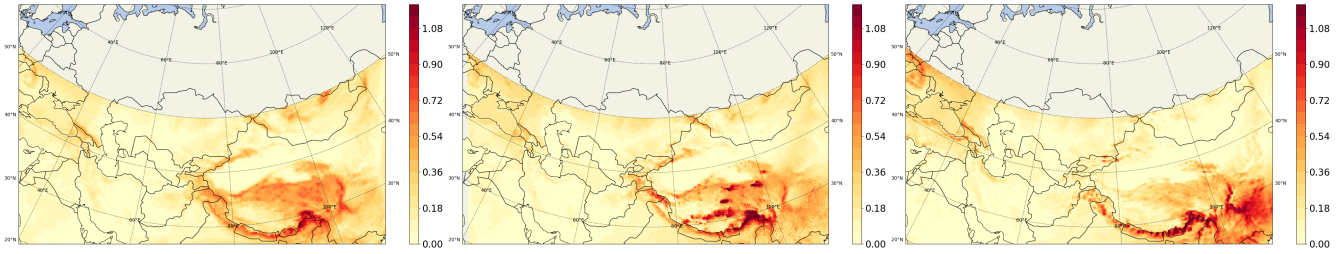
respect to ERAInterim (Fig. 2, panels *e,f,g*). However, even in those regions, e.g., in the South and Southeastern parts of the domain, positive values of A calculated with respect to CHIRPS are observed for the RCMs, but not for the GCMs. This might be because RCMs capture more detailed convective-forced precipitation than GCMs and we already removed 10 grid points of CCLM data at the lateral boundaries as the so-called buffer zone for our analysis where the model is forced to follow the forcing GCM. The precipitation pattern over the South and Southeast of the domain is mostly connected to the monsoon, and RCMs better fit the CHRIPS than ERAInterim (Fig. reffig:2.a−d).

Added value of CCLM driven by MPI-ESM1-2-HR1-2

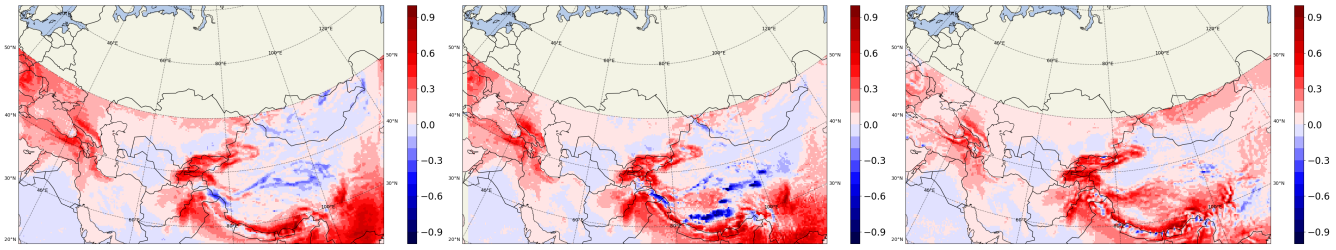
We showed that COSMO-CLM could reduce the large-scale bias of its driving reanalysis for daily precipitation, especially over areas with a complex topography and the Asian monsoon region. Here, we calculate the added value of the CCLM simulations driven by MPI-ESM1-2-HR for 1985-2014. It can be seen in figure 3.a that the driving GCM shows a more negligible bias than the ERAInterim over Tajikistan and Kyrgyzstan. According to Déqué et al. (2007), the GCM bias is one of the most important sources of uncertainty in the RCM's regional climate projection, and the smaller $D_{MPI-ESM1-2-HR2}$ compared to $D_{ERAInterim}$ over Tajikistan and Kyrgyzstan might increase the skill of the final regional projections (under the assumption that the model bias remains conserved under other radiative forcings). The added value of CCLM driven by MPI-ESM1-2-HR shows small values over those areas compared to the simulation driven by ERAInterim. However, the bias over the South and the north band of the Tibetan Plateau and the East and West of the domain is reduced substantially. Our analysis of the two driving datasets (ERAInterim and MPI-ESM1-2-HR) tends to confirm the findings of the Sørland et al. (2018), at least of the total precipitation PDFs, that the biases of the GCM-RCM chain are not additive and not independent. For example, in almost all regions with high values of yearly precipitation, where GCM has a slight bias, the RCM does not present higher biases or vice versa.



(a) D(ERAInterim)

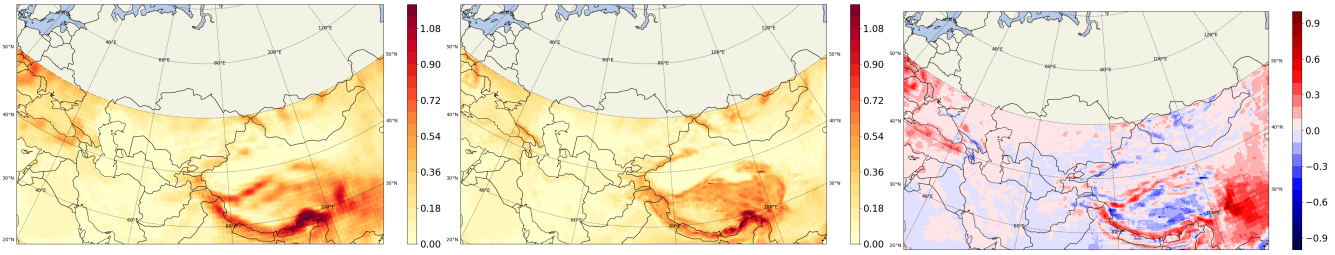


(b) D(CCLM-ERAInterim) (c) D(GERICS-REMO2015-ERAInterim) (d) D(RMIB-UGent-ALARO-0-ERAInterim)



(e) A_i CCLM (f) A_i GERICS-REMO2015 (g) A_i RMIB-UGent-ALARO-0

Figure 2: Relative probability difference (D) for ERAInterim (a), CCLM-ERAInterim (b), GERICS-REMO2015-ERAInterim (c), ERAInterim-RMIB-UGent-ALARO-0 (d) and added value (A_i) for the (e) CCLM-ERAInterim, (f) GERICS-REMO2015-ERAInterim and (g) ERAInterim-RMIB-UGent-ALARO-0 compared to the CHIRPS at 0.22° horizontal resolution.



(a) D(MPI-ESM1-2-HR) (b) D(CCLM-MPI-ESM1-2-HR) (c) A_i

Figure 3: Relative probability difference (D) for MPI-ESM1-2-HR (a) and CCLM-MPI-ESM1-2-HR (b), and added value (A_i for the CCLM-MPI-ESM1-2-HR compared to the CHIRPS at 0.22° horizontal resolution.

this should be f, i, l

Done!

P9: This doesn't make any sense when reading. So was the sentence before the higlight "areas where ..." about summer? And it looks like in winter there are actually big magenta blobs, especially for GERICS-REMO2015 (also for annual). You should also analyse this.

We modified the whole two paragraphs in the new version of the manuscript according to your comment.

P9: I'm guessing this comes from Fig.5? If yes it should be referenced. And again you use "significantly" which is a very strong statement and requires statistical testing. Use a synonym. Above when referencing you use Figs, should be Figs.4a-c for consistency... Also you're referecning the GCM here while talking about the RCMs. I'd say remove the Fig reference and this should come before the previous sentence talking about bias...This should come before and is not on a border, so boundary effects don't seem to be a good explanation here What type of biases because one seems to be positive (in green for tibetan) and the other negative (in magenta)

We modified the whole paragraph accordingly. We made it clear that the RCM's performance near the boundaries is influenced by the constraints of the driving GCM, reducing its ability to improve results in those regions.

P11: No need to repeat this every time

Done!

Unless you actually compare the speed of dynamical DS versus ML that's an empty statement. Same for computational cost.

Done!

P11: What's the time period of this analysis? It's also not mentioned in the Figure...

Although the dataset covers a specific time period, the exact dates are not critical for this analysis because the data was randomly shuffled at the pair level before training, validation, and testing. This shuffling ensures that the model is exposed to diverse conditions without introducing temporal or spatial bias. Therefore, the time period is not relevant to evaluating the model's performance, as the focus is on ensuring generalization through randomized data exposure. For the distribution, 68,141 days (60%) were used for training, 22,714 days (20%) for validation, and 22,714 days (20%) for independent testing, regardless of the specific dates.

While the dataset was shuffled for the original model training, validation, and testing to avoid temporal or spatial biases, ensuring the model's generalization performance, we took a different approach when evaluating the CNN with a new GCM later in the study. For this evaluation, the dataset was not shuffled to maintain the temporal structure and calculate the correlation, as the goal was to assess the model's performance with respect to the original time series of the new GCM. This allowed us to evaluate the model's ability to capture temporal patterns and validate the results accordingly.

We clarified that in the Figure.

P11:it's relative to the RCM

True, we modified the text.

And what happens at the places where the constrained models reflect the GCM grid? Why is it negative there for the constrained models and not for the unconstrained?

A very good point!

In Figure 4 of this answer we focus on the areas mentioned by you: green areas indicate improvement (reduction in error, i.e., positive added value) while magenta areas show worsening performance (negative added value). The blue lines highlight the regions where the GCM has maximum errors.

Interestingly, along these blue lines, the CNN shows no significant improvement compared to the GCM, implying that the CNN struggles to reduce the large errors where the GCM performs poorly. This behavior likely results from the constraints placed on the CNN—such as mass conservation or the preservation of physical quantities like precipitation—which limit its ability to freely adjust the outputs in these areas.

However, in the regions immediately surrounding these blue lines, we observe a mix of positive and negative added values, with green areas indicating that the CNN is successfully reducing errors, while the magenta areas suggest over- or under-compensation. This pattern may arise because, while the CNN is attempting to reduce the GCM's

large errors, it is constrained by the overall balance of physical variables (such as the mass of precipitation), causing it to redistribute errors into nearby regions.

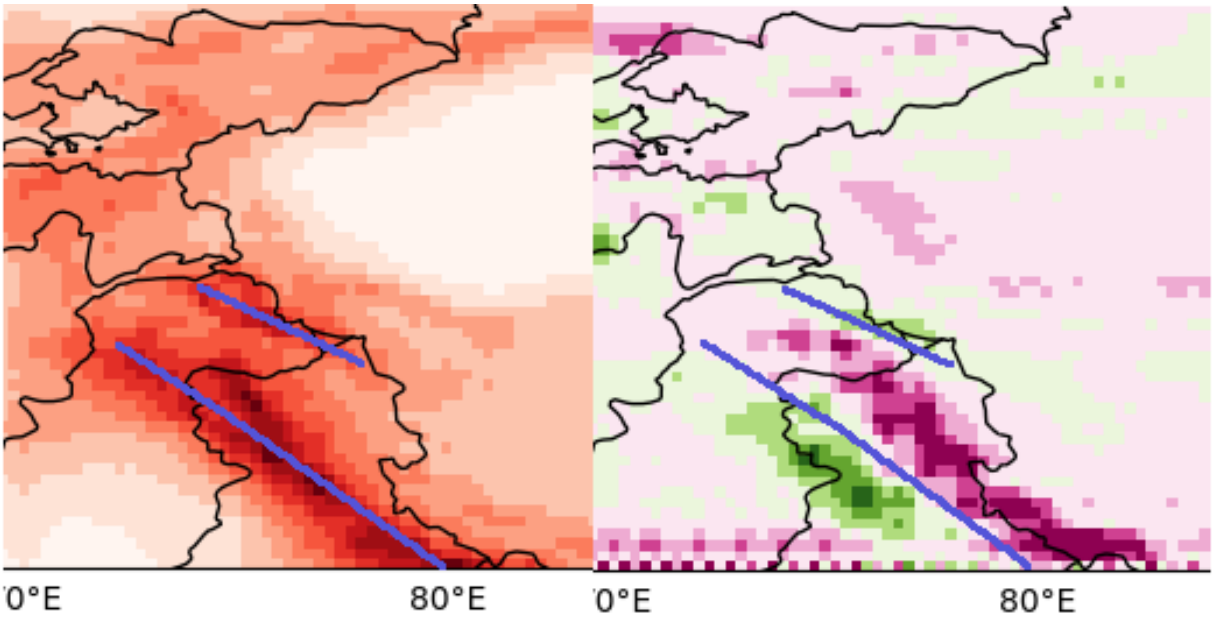


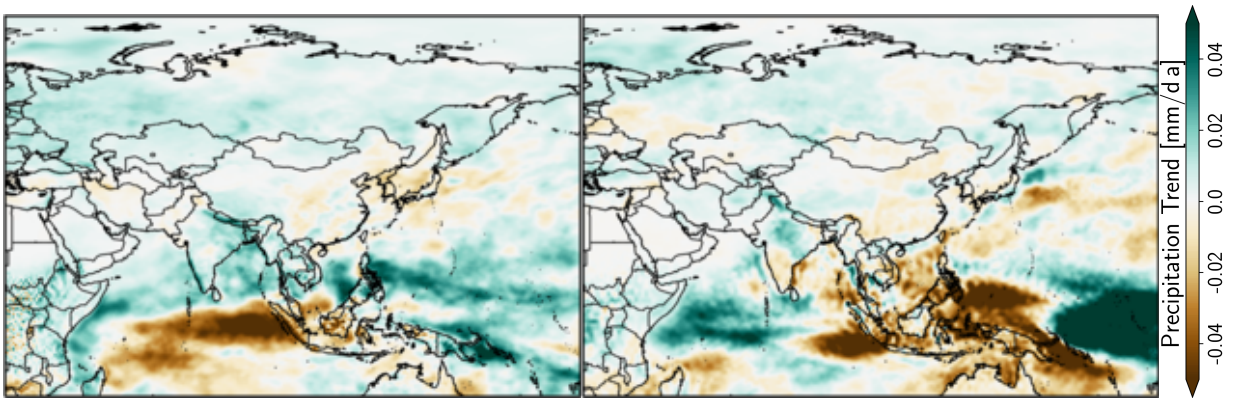
Figure 4: Left panel: GCM error distribution over the domain. Right panel: CNN error reduction compared to the GCM. Green areas represent regions where the CNN reduced errors (positive added value), while magenta areas show regions where the CNN increased errors (negative added value). The blue lines show places where the GCM shows maximum errors.

Fig. for consistency... how much? This should be in the text... which is

Done!

P12: You chose this GCM so it's not remarkable that they are very similar. You could have also chosen one that was different.

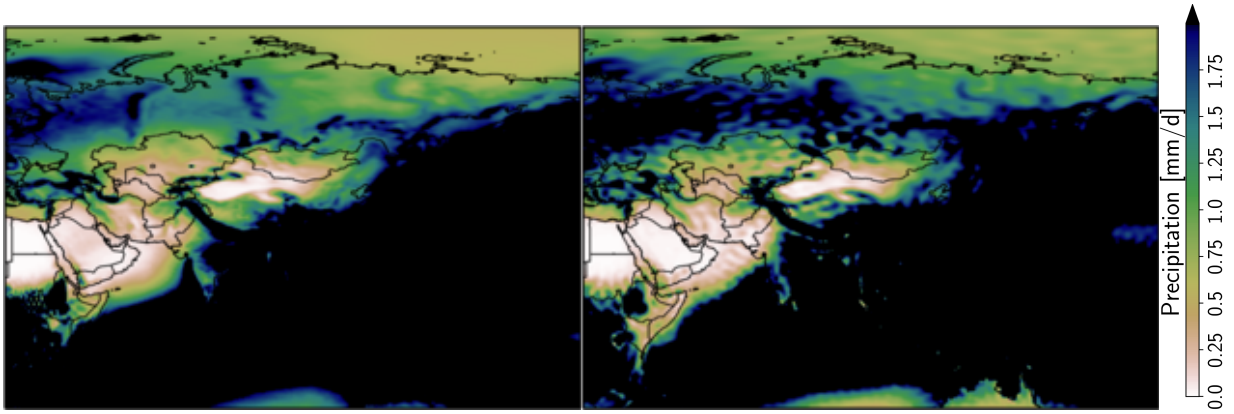
Yes, we chose this GCM, but it was not cherry-picked. This model is one of the 10 selected models from the CMIP6 ensemble for the ISIMIP project. These 10 models were specifically chosen because they represent diverse physical processes and model codes across all components of the Earth system, ensuring a wide spread of the CMIP6 models. The selection was designed to cover a broad range of potential outcomes, which is precisely why we used this model. For example Figure 5 of this answer shows how the precipitation annual trend for 1981-2014 are different in those models.



(a) EC-Earth3-Veg (b) MPI-ESM1-2-HR (c)

Figure 5: Asia Precipitation annual trend 1981-2014 single models of CMIP6

We also show the absolute values of pr then here in Figure 6.



(a) EC-Earth3-Veg (b) MPI-ESM1-2-HR (c)

Figure 6: Asia Precipitation annual absolute 1981-2014 single models of CMIP6

But we remove the word "remarkable" in the new version of the manuscript. We agree that the generalization ability depends on whether the new dataset is taken from the same distribution of values. They were driven by the same forcing which are typical for any SSP scenario setups.

P12: Yes but this was also not the case for the NoCL under the normal GCM so unsurprising.

Done!

P12: This should come before in the paragraph when the emulator is first mentioned.

Done!

P12: has improved compared to what? And give numbers..

We added the reference and the numbers were already presented in the next sentence.

P12: This should information come way before in the beginning of the paragraph as it's very important.. .

We added that information to the beginning of the subsection and modified the text.

P12: has a lower MAE than the GCM

Done!

P12: ressembles

Done!

P12: but how similar is this unseen forcing to the one the emulator has been trained on? This information is missing.

Figure 7 shows the differences in total precipitation between the SSP scenarios for the MPI−ESM1−2−HR model. As can be seen there are differences in the climatology of SSP370 and other SSPs.

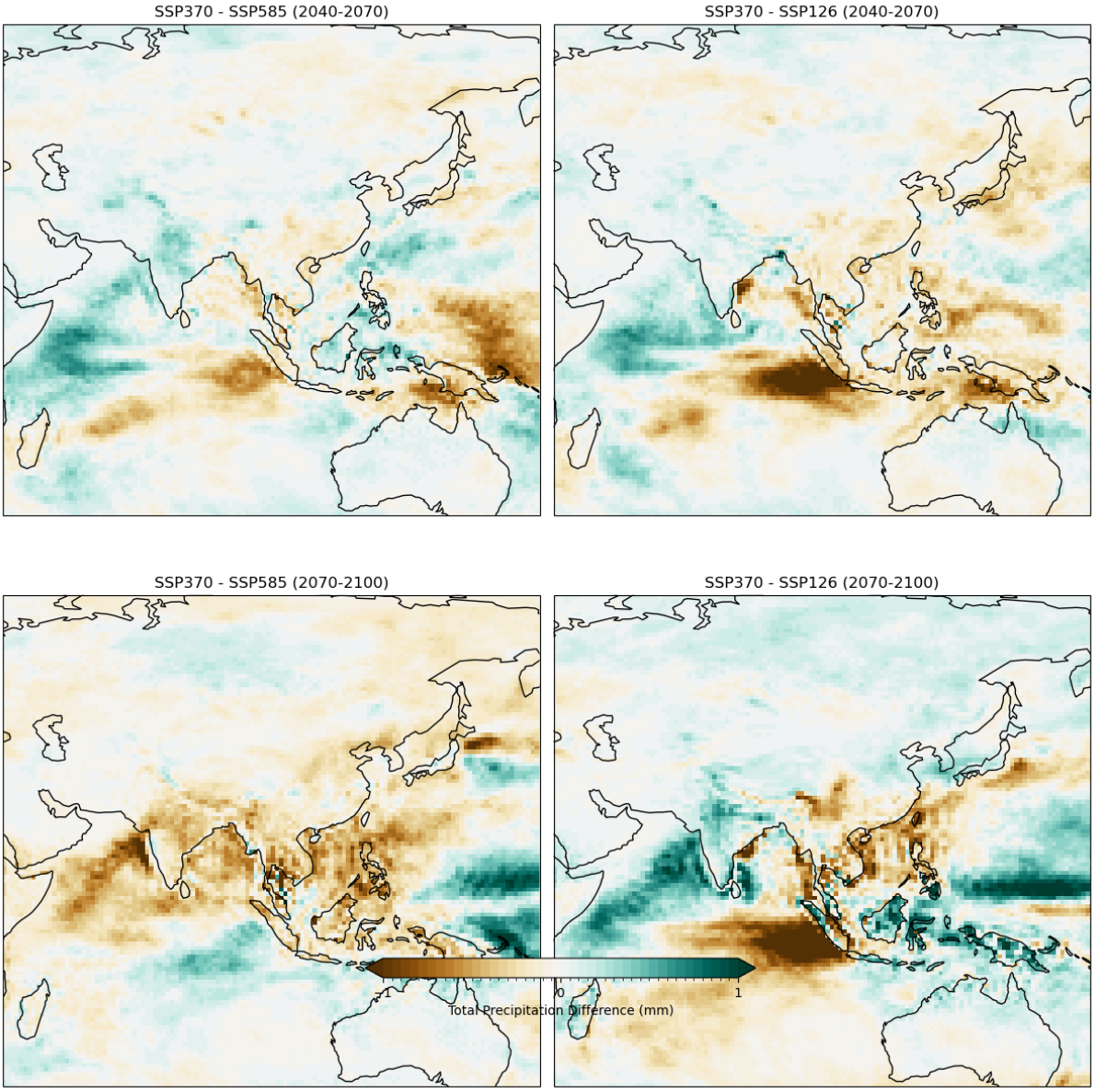


Figure 7: Differences in total precipitation between the SSP scenarios for the MPI−ESM1−2−HR model. The colorbar indicates the total precipitation difference in millimeters.

P12: No not really, it depends on how similar this new SSP is to the ones it was trained on, this statement is coming on way too strong. Maybe rewrite into "shows promise for learning and reproducing the [...].

We changed the text to what you have suggested.

P13: of RCMs

Done!

P13: of RCMs

Done!

P13: Yes ok but you didn't do that so it feels weird to add this here. Except if you reformulate it into something that should be done in further work.

We reformulated that to : "In future work, it would be valuable to follow the approach suggested by Volosciuk et al. (2017), where downscaling outputs are evaluated at coarser resolutions. This would allow for a deeper understanding of how downscaling methods introduce or fail to correct biases, which can vary significantly across spatial scales. By conducting evaluations on a coarser grid, we can better distinguish between the inherent biases of the model and those introduced by the downscaling process, providing important insights into the limitations and strengths of downscaling techniques in representing climatic variables across different scales."

P13: No I really don't agree with this as Paleo climate experiments will be very different. Yes it learns a relationship but only for climate variables from the training set. So if you suddenly give it extreme values like for paleo climate it might not work and needs to be tested before making this statement. Guillaume Jouvet's CNN was a very different architecture and was trained on paleo climate directly not transferred like you propose. So this sentence feels wrong here.

We removed those two sentences.

P14: No again this statement is too strong. Maybe it "shows promise" but as I said before you didn't test the similarity of this unseen scenario to the training. Maybe it just fits nicely in there but it might not work for another more extreme scenario.

We answered this and showed that the two models are not very similar. However, we modified the text in a way that it is not that strong.

P21 : This figure ideally should also contain the time frame to know how many samples and from what time we're looking at.

Done!

P23 : MAE defined if you define added value.

Done.

1.2 Other comments

Overall results: you use a lot of statements that need values from the figures to back them up; otherwise, they seem superficial. For example, if you say the "correlation of the CNN is higher than the RCM (here, include a value from the figures that supports this claim)." Also remove empty adjectives like "strikingly" or "remarkably" as they are very subjective and results should be based on numbers.

As already answered. We revised all those "superficial" adjectives and put numbers and analysis.

3.1 Added value of CCLM driven by ERAInterim: needs some shuffling of sentences, I would also present the MAE and bias results together instead of two separate paragraphs seeing how they are very similar. You can do this by using "this is also reflected in the bias blabla" somewhere.

Done in the new version of the manuscript.

4 CCLM emulator using a CNN: It needs values to support claims, and an analysis of the difference in negative AV for constrained CNN versus positive AV for unconstrained CNN would be nice (I'm thinking here of the magenta blob in the lower part of the image).

We discussed that before in this answer and added the analysis (Fig. 4 of this answer).

4.1 Applying the CNN to a different GCM: The new GCM you chose is very similar to the one the model is trained on. This should not be presented as a surprise but as a design choice that will impact the results, seeing how generalization is easier for the model then. I think that, in your case, generalization works pretty well, seeing how the map you produce is very similar to what you showed in Fig. However, the text should mention that for a more different GCM, it might work less well. The same applies to the new SSP; you should mention somewhere how similar this SSP is to what is in the training set. Otherwise, it's hard to evaluate the generalization ability. And if your model didn't work well on new GCM/SSP that is very different from training, it would not be a surprise or a downside of your model; that's just ML, but it needs to be transparent in the text, and that generalization ability depends on whether the new dataset is taken from the same distribution of values as the training and otherwise it needs special ML techniques like transfer learning.

In the revised manuscript, we have clarified that the new GCM we chose is not very similar to MPI-ESM. We included a comparison of the trends and climatology of these models and added a discussion on the ISIMIP model selection. Following your suggestion, we have removed any strong conclusions from the results and highlighted that

the performance of the CNN is indeed tied to the GCM it was trained on, as all models use the same SSP forcing. Additionally, we now emphasize the potential limitations of generalization when applying the model to more distinct GCMs or SSPs.

Discussion: Following what I said above, I really disagree with the statement that your model could be applied to paleoclimate, as the distribution of paleo climate variables is probably very different from your training set. I would remove all this, also the mention of Jouvet's work as it is very different from your case, and just mention that your CNN shows should work for RCMs from the same GCM as the training one, or for similar GCMs as you have shown in your results.

Following your suggestion, we removed those parts from the new version of the manuscript.

Generally, you show interesting results, but with your ML model, you should be very careful about the promises you make of its use and stay realistic, seeing how it is very limited to situations similar to its training (not because you did something wrong, but just because of ML in this setup).

We completely agree with your point, and in the revised manuscript, we have carefully addressed this by ensuring that our promises regarding the use of the ML model remain realistic. We have taken care to present the model's limitations and emphasize its applicability to situations similar to its training. .

On behalf of all authors,

Bijan Fallah

References

Bengio, Y.: Practical recommendations for gradient-based training of deep architectures, in: Neural networks: Tricks of the trade: Second edition, pp. 437–478, Springer, 2012.

Ciarlo, J. M., Coppola, E., Fantini, A., Giorgi, F., Gao, X., Tong, Y., Glazer, R. H., Torres Alavez, J. A., Sines, T., Pichelli, E., et al.: A new spatially distributed added value index for regional climate models: the EURO-CORDEX and the CORDEX-CORE highest resolution ensembles, Climate Dynamics, 57, 1403–1424, 2021.

Déqué, M., Rowell, D., Lüthi, D., Giorgi, F., Christensen, J., Rockel, B., Jacob, D., Kjellström, E., De Castro, M., and van den Hurk, B.: An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections, Climatic Change, 81, 53–70, 2007.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., et al.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes, Scientific data, 2, 1–21, 2015.

Goodfellow, I.: Deep learning, 2016.

Harder, P., Hernandez-Garcia, A., Ramesh, V., Yang, Q., Sattegeri, P., Szwarcman, D., Watson, C., and Rolnick, D.: Hard-Constrained Deep Learning for Climate Downscaling, Journal of Machine Learning Research, 24, 1–40, 2023.

Sørland, S. L., Schär, C., Lüthi, D., and Kjellström, E.: Bias patterns and climate change signals in GCM-RCM model chains, Environmental Research Letters, 13, 074 017, 2018.

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.