Review of ‘Climate Model Downscaling in Central Asia: A Dynamical and a Neural Network Approach’

by

B. Fallah, C. Menz, E. Russo, P. Harder, P. Hoffmann, I. Divodets, and F.F. Hattermann

Recommendation: major revision

The study presents results from two downscaling methods for GCM precipitation. 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.

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.

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.

Some of the specific problems are listed below.

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? 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.

Line 11: ‘we downscale CCLM’ is not correctly phrased

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.

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.

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.

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

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'.

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?

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.

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?

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.

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?

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.

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.

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.

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.

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

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.

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'.

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.

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.

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.

Line 322: Project reference is missing

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.

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

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