Using Multi-Head Attention Deep Neural Network for Bias Correction and Downscaling for Daily Rainfall Pattern of a Subtropical Island

General comments

In this work, authors propose an attention-based deep learning model for the downscaling and bias correction of precipitation over Taiwan. By relying on precipitation and 10-m wind at a coarse scale (25 km), the model is able to generate the downscaled field at a regional scale (5 km). To achieve this, they use a reanalysis dataset as the input (predictor) and an observational dataset as the output (predictand). They also integrate orographic data into the model to better represent local precipitation affected by orographic phenomena. By comparing the proposed model against a baseline bias-correction algorithm, they demonstrate the improvement of the proposed model across a wide set of metrics, including those characterizing the mean, precipitation extremes, and interannual variability, as well as specific extreme precipitation events. As a result, the authors plan to continue working on applying this model to climate models to generate regional-scale projections for future scenarios.

First of all, I recommend that the authors review the use of the English in the manuscript, particularly with respect to some sentence structures and inconsistent verbal tenses. Also, I believe that the organization of the manuscript and presentation of the data and results need some improvement. I leave some specific comments regarding this in the *Specific Comments* section.

From my perspective, there are some misconceptions regarding important concepts in the field of deep learning for statistical downscaling. For instance, in the introduction (L36-49), authors define statistical downscaling as follows:

“Statistical downscaling, conversely, constructs empirical relationships between coarse-resolution variables from GCMs and local-scale surface variables”

However, as defined in [1]:

“In statistical downscaling, empirical links between the large-scale and local-scale climate are identified and applied to climate model output.”

Whether these links are constructed using simulations/models (e.g., GCMs) or observations/reanalysis as input characterizes the Model Output Statistics (MOS) and Perfect Prognosis (PP) approaches, respectively. Thus, the authors are wrongly defining statistical downscaling as MOS, since the latter is a subset of the former.

In the next paragraph, authors define Super Resolution (SR) techniques. As they correctly mention, these techniques have gained popularity due to advances in the deep learning field. However, the main drawback of SR models is that they generally rely on the upscaled surface variable or an equivalent from a different observational/reanalysis dataset (e.g., ERA5). One could relate this approach to Perfect Prognosis (PP); however, in PP, models are not built with surface variables as predictors (among other assumptions), as these are not well reproduced by GCMs (this would lead to biased projections when downscaling the GCM). This is especially relevant for variables with a heavy local-influence such as precipitation, which may substantially differ between observations and GCMs. Instead, PP relies on large-scale synoptic variables, which are properly simulated by GCMs.

SR techniques are also inappropriate for Model Output Statistics (MOS) in the climate context, as these techniques assume a day-to-day correspondence between the simulated data (GCM in this case) and observations, which is not fulfilled in this context. This is why, in the climate context, MOS is performed distributionally (e.g., the BCSD technique authors use as baseline).

These considerations align with the downscaling model proposed in this paper. In this work, authors downscale precipitation (5 km) by relying on precipitation and 10-m wind data from ERA5 (25 km). Since this framework does not align neither PP or MOS assumptions, I recommend removing all mentions of applications of the proposed deep learning model to GCMs and instead framing it within the context of SR.

I would also like to highlight that the deep learning model proposed in this paper is based on the one developed in [2], with the addition of the attention mechanism and the 10-m wind covariate. However, the experimental framework of these two works is very similar. In Section 3.6.1, the effect of including the 10-m wind variable is addressed. However, to justify the use of attention-based layers, the proposed model should be compared to that of [2], or at least to some other CNN model previously discussed in the literature (e.g., [3]).

Specific comments

L23*"[...] between synoptic and local circulations influenced by topography."***:** The proposed model is fed precipitation and 10-m wind as predictors, which are considered surface variables rather than synoptic variables.

L36-49: I suggest rewriting this paragraph following my general comment on the definition of statistical downscaling.

L43 *"[...] the Model Output Statistics (MOS) [...] requiring no prior knowledge for the of predictors or regions."*: I believe that such a strong statement requires a reference.

L64-69: I would recommend trying to better categorize these papers based on my comments on the definition of statistical downscaling in the general comments section.

L91-93/153-155 *”we have refined the training and validation procedures by opting for partitioning based on consecutive years”*: This practice has been already used in numerous downscaling works, such as [4, 5], among others. I believe these should be mentioned.

L101: Are environmental *representations in GCMs* the same as the synoptic scale?

L101-102 *“Our methodology is tailored for future climate downscaling rather than nowcasting”*: As I mention in the general comments, I believe authors need to reconsider the scope of this work regarding future applications.

L120-124: If, as authors mention, they want to reproduce the interplay between synoptic weather patterns and topography you need to rely on synoptic variables, but the developed model relies on surface ones.

L149-151: What is the purpose of the validation set in this study? Generally, this set is used to find the best set of hyperparameters, performing the final evaluation in the test set. May it be used here for the early-stopping process?

L157-158 *”These steps encompass a log1p transformation to adjust for the skewness in the distribution of the data values, particularly beneficial for precipitation data”:* Is this transformation applied to all variables or just to the precipitation? It is not clear.

L158-159 *”Moreover, we normalize various data variables to ensure consistency across the dataset: precipitation, temperature, and humidity data [...]”:* In L124 you say: “*To capture these complex interactions, our model, EDA, incorporates ERA5-derived daily aggregated rainfall and 10-meter height wind data as inputs”.* What variables did you really use as predictors? There are some contradictions in the manuscript.

L186-187 *“[...] As for the downscaling process, the intermediate outputs from the encoder are transitioned to the decoder, [...]”*: I do not understand this. Following Figure 2, it seems that the intermediate outputs of the encoder are not transitioned to the decoder, instead the final encoder’s output is passed to the decoder. I believe this needs further clarification.

L192-197: I understand that the use of the pixel-shuffle layer for upscaling is inspired by the Enhanced Super-Resolution CNN, but what about the Super-Resolution Using Deep Convolutional Networks you mention in the paragraph before? Is any element of this model used in the model proposed in this work?

L202-203: Regarding the WMSE loss function, is this the first time such a metric is used or did you draw inspiration from some other work? Also, I think it is necessary to specify what value of γ do you set. If you choose γ=0 you get the standard MSE, otherwise if γ=1 you get the MSE weighted by the true precipitation value. The latter can be problematic for non-precipitation cases, as these would not contribute to the loss function. This is problematic both for the raw precipitation value as for the lop1p-transformed (non-precipitation would still correspond to 0).

L205-208: I believe readers should be provided with more details regarding the pre-training phase of the encoder. If I understand correctly, the encoder is initially pre-trained on low-resolution data. However, it is not explicitly defined what this data corresponds to. Is it a low-resolution version of the target data? Additionally, it's unclear if the same loss function is used during this pre-training phase.

L206-207 *"[...] the encoder is frozen, and the encoder is trained on high-resolution data, [...]"*: I guess this is a typo, as when you freeze the encoder you train the decoder.

Figure2: This figure indicates that the input data have a channel size of 4, which does not align with the variables introduced in Section 2.1 (see my L158-159 correction).

L230: I couldn’t access Lin et al., 2023, the paper in which the baseline method is based on.

L257: What set of additional climate indices are you referring to?

L290 *“with a bias residual of less than 2 mm/day (Fig. 3b, 3c)”*: You mention the bias residual, but there are no biases plotted for the mean rainfall. I recommend either adding an additional row to this figure displaying the climatology of the observational dataset or following the format of other figures (e.g., Fig 4, Fig 5, Fig 6) and including the bias in the plot.

Caption of Figure 4, 5, 6, 9, 10: Why does the test period start on such a specific day (2017/12/13)? Is this a typo, or does the test period actually begin on this specific day?

L323 *“Figure 5 displays the spatial distribution of days experiencing rainfall exceeding 10 mm (RX10m) [..]”*: Does RX10m represent the number of days or the amount of precipitation for days with 10m? It is not clear in the text, as in L255 you say *annual account* but then in Figure 5 the colorbar has the label *mm/day*. This needs further clarification.

Figure 6: The label of the colorbar for the biases should be days instead of mm/day .

L340-341 *“During the summer’s wet seasons, BCSD consistently overestimates CDD throughout Taiwan”*: If I'm not mistaken, BCSD underestimates the CDD instead of overestimating it.

L341-342 *”In fall, BCSD’s predictions overestimate CDD in northeastern Taiwan and similarly overestimate CDD in western Taiwan”***:** In northeastern Taiwan, BCSD underestimates CDD instead of overestimating it. I suggest reviewing this paragraph thoroughly, as there may be some misunderstandings regarding the under/overestimation of the CDD metric.

L344-345 *“[EDA] tends to underestimate CDD during fall, indicating a nuanced yet imperfect prediction capability for dry periods throughout the seasons.”*: Could this also be attributed to the WMSE loss function? Without specific details about the chosen γ in the manuscript, it's uncertain. However, a high value of this hyperparameter might cause the model to overestimate the precipitation amount, consequently resulting in an underestimation of the CDD.

L347-349 *“The challenge in accurately modelling this season may stem from the substantial contribution of typhoon-related rainfall, which due to its somewhat stochastic nature compared with other seasons, complicates the precise prediction of rainfall distribution”*: Some of the differences shown in Table 4 are quite small, which makes drawing conclusions difficult. Have you considered, instead of presenting results for a single model, training the model multiple times and reporting the mean metric across these runs along with a measure of variability? This approach would facilitate the assessment of whether the EDA model differs significantly from the BCSD.

Section 3.3: Why is the zero (no precipitation) not represented in Figure 7? MSE-based deep learning models often tend to fit the mean, which can be problematic for a variable like precipitation. Including visualization of the 0 mm category would be beneficial. Additionally, have you considered using other visualization methods? For example, a histogram with the y-axis in logarithmic scale could provide better visualization of extreme values.

L378 *“Our analysis, spanning from 2015 to 2020, includes both testing and validation phases”*: I believe that using the validation period to assess the final accuracy of a deep learning model may lead to a biased assessment of its accuracy, as this period is typically utilized to search for the optimal configuration. If the validation period has not been used to tune the model, then it should be considered part of the test period.

L436-437 *“Figure 10a reveals that mean rainfall is significantly underestimated by EDA_PR when relying solely on ERA5 rainfall data as input [...]”*: This is the first time the bias of the mean is shown. In Figure 3, the mean is displayed, but not the bias with respect to the target data. Would it be possible to also show the bias in the mean for the EDA model? This would facilitate comparing EDA_PR with EDA in terms of this metric.

L464 *“[...] adopting reinforcement neural networks, such as generative adversarial networks, [...]”:* As far as I know, reinforcement and generative learning are two distinct paradigms. Therefore, a generative adversarial network is not a reinforcement neural network.

Technical corrections

L29-30: I suggest including [6] (Section 1.5.3) as reference for GCMs/ESMs.

L104-109: Instead of Sections, the text wrongly refers to these as Sessions.

L144 *“[...] as outlined by X”*: This appears to be a typo.

L176: This should refer to Figure 2, not Figure 3.

L166 *“[...] an encoder and an encoder”*: This should be *“an encoder and a decoder”*

L220 *“[...] for its capability to maintain the mean percentile of data distribution efficiently, [...]”*: I believe this sentence could be improved, for instance by simply saying: “*for its capability to reproduce the mean”.*

L241-243: This paragraph needs some rephrasing for better understanding.

L253: SDII stands for Simple Daily Intensity Index, not Simple Precipitation Intensity Index).

L255: Following [7], this metric should be denoted as R10mm, not RX10mm.

L272 *“[...](depicted by red and purple lines in Fig.1b)[...]”*: Are you referring to the test and validation precipitation series? If so, you could simply mention that such extreme precipitation events occur during these periods.

L276 *“[...](Fig.1b)[...]”*: I believe you are not referring to Fig. 1b here, but rather Fig. 1c.

L282-286: You are referencing subplots that are not defined (e.g., Fig 1e or Fig 1f).

Figure 8: This figure requires some improvements in terms of visual appearance, such as addressing issues like no data for 2014 being plotted on the x-axis and the presence of empty space in half of the plot. Additionally, it would be beneficial for the reader to use consistent colors to represent the same models across Figures 7 and 8.

References:

[1] Maraun, D., & Widmann, M. (2018). Statistical downscaling and bias correction for climate research. Cambridge University Press.

[2] Chiang, C. H., Huang, Z. H., Liu, L., Liang, H. C., Wang, Y. C., Tseng, W. L., ... & Wang, K. C. (2024). Climate Downscaling: A Deep-Learning Based Super-resolution Model of Precipitation Data with Attention Block and Skip Connections. arXiv preprint arXiv:2403.17847.

[3] Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2017, August). Deepsd: Generating high resolution climate change projections through single image super-resolution. In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining (pp. 1663-1672).

[4] Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. Geoscientific Model Development, 13(4), 2109-2124.

[5] Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. Journal of Advances in Modeling Earth Systems, 14(10), e2022MS003120.

[6] Chen, D., M. Rojas, B.H. Samset, K. Cobb, A. Diongue Niang, P. Edwards, S. Emori, S.H. Faria, E. Hawkins, P. Hope, P. Huybrechts, M. Meinshausen, S.K. Mustafa, G.-K. Plattner, and A.-M. Tréguier, 2021: Framing, Context, and Methods. InClimate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change[Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 147–286, doi:10.1017/9781009157896.003.

[7] https://etccdi.pacificclimate.org/list_27_indices.shtml