Review of the manuscript:

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

In this paper, the authors proposed a deep learning model called EDA to downscale the daily rainfall precipitation from ERA5 with $0.25^{\circ} \times 0.25^{\circ}$ to 5-km observed rainfall data over Taiwan. The paper can be seen as an improved version of Chiang et al. (2024) where the authors replaced the channel and spatial attentions (Woo et al. 2018) by a self-attention from Vaswani et al. (2017). In addition, 10-m wind components and high-resolution topographic data were used as auxiliary information and the loss function was replaced by a weighted mean squared error (WMSE). The authors show that the model can perform a downscaling and bias correction from ERA5 to match the high-resolution of TCCIP observational data.

General comments

First, since the manuscript introduces an improved version of an existing model from Chiang et al. (2024), it should provide ablation studies about the newly introduced model components. In my opinion, this should definitely include ablation studies about the:

- proposed loss function.
- self-attention layers in comparison with the one from Chiang et al. (2024) and with CNN.
- encoder block in comparison with the standard block design from Transformer by Vaswani et al. (2017).
- training procedure with two phases (see lines 205-208).
- sensitivity studies about auxiliary information i.e., topographic data and temperature/humidity.

I also find it confusing to name the attention module *"multi-head Attention layers for auxiliary channels (EDA)".* As mentioned in lines 170-172: *"...This capability is pivotal for our model, allowing it to simultaneously process the entire dataset and enabling each grid point to evaluate its relationship with all others..."* the layer performs attentions across the spatial domain and not between the channels i.e., the attention matrix will be 126×126 in this case compared to 4×4 in case of using 4 auxiliary predictors without increasing the dimensionality (see Fig. 2).

Second, I think the manuscript should indicate clearly what ERA5 variables were used in the study. Throughout the manuscript, it is stated that 10-m wind components were used as auxiliary predictors while line 159 introduces suddenly two more predictors (temperature and humidity). Figure 2 implies also that 4 input variables were used.

Third, it looks like there is a misunderstanding about the applications of this study. The authors wrote in lines 101-102: *"...Our methodology is tailored for future climate* *downscaling rather than nowcasting…”* and in lines 514-515: *“…We intend to initiate this expansion by applying selected CMIP6 models for climate downscaling, aiming to generate precise local-scale climate projections for Taiwan, by harnessing data from East Asia or potentially global reanalysis…”* Reanalysis data and climate projections are different. I don’t think the experimental framework is designed in a way that the model can be used to downscale climate projections since training was done on ERA5 reanalysis.

Lines 162-163 *“These improvements collectively contribute to an increase in the model's computational efficiency and predictive accuracy”:* What do the improvements refer to? Did the manuscript prove that the referred improvements increased computational efficiency and accuracy? If yes, where in the manuscript?

Lines 205-208 *“…The training regime is executed in a supervised manner, with an initial focus on training the encoder using low-resolution observational data. Subsequent to this phase, the encoder is frozen, and the encoder is trained on high-resolution data, a strategy designed to fine-tune the model's ability to perform accurate downscaling and bias correction…”*: Where was it proven in the manuscript that the newly introduced strategy improves the accuracy of downscaling and bias correction?

Subsection 2.3: Model Structure: In my opinion and since the manuscript is submitted to (GMD), this section lacks essential implementation details. For instance, what is the number of layers in the encoder (*N* in Fig. 2)? What are the kernel sizes, strides and paddings for CNN? How is the positional embedding implemented? is it learnable or based on a fixed sinusoidal embedding? What kind of activation function is used? How are the topographic data normalized? What kind of normalization layers is used in the model? is it LayerNorm Ba et al. (2016)? What is the drop out ratio? How is the Up-sampling done? What does subpixel mean? Is it interpolation or ConvTranspose? How many attention-heads are being used? How big is the model and how many parameters does it have? In addition, self-attention should be described technically.

Implementation details: I found some differences between the submitted code on Zenodo (Chiang, 2024a) and the parameters mentioned in the manuscript. For instance, γ is set to zero in the code, early stopping was set to 20 epochs instead of 60 and learning rate was set to 0.00005 instead of 0.0001 in the manuscript.

Subsection 2.4 Baseline Downscaling Methods: Since the experimental framework and data are very similar to the ones used in Chiang et al. (2024), I think the deep learning models from Chiang et al. (2024) should be included as baselines.

I would also like to emphasize that the manuscript should acknowledge the limitations i.e., the computational requirements for self-attention grow quadratically with resolution or that the model was trained with a biased corrected Reanalysis (ERA5) as input and not with a climate projection which makes it unapplicable to downscale climate projections.

I also suggest that the authors rearrange the manuscript in a way that they move the figures and tables from the appendix and put them in context within the main body of the manuscript. Furthermore, the manuscript needs to be checked for references to tables and figures inside the text and the usage of language, especially regarding the conjunctions and connecting words/sentences (see specific and technical comments).

Specific and technical comments

Lines 10-13 *"Abstract. This study investigates the capability of a deep learning approach, employing a multi-head attention mechanism within a deep neural network (DNN) framework, aimed at refining the bias correction and downscaling process for the fifth generation European Centre for Medium-Range Weather Forecasts reanalysis rainfall datasets to provide local-scale daily rainfall data across Taiwan, a mountainous subtropical island…"*: I think this is a very long sentence. Please split the sentences and ideas in the manuscript.

Line 87: grammar, mean squared error (MSE).

Lines 96-97 *"…incorporating surface winds and topography as auxiliary datasets…":* did you only used wind components as predictors?

Line 140: What kind of topographic data is used? Slope, DEM, …?

Lines 29-30: Please add references for GCM and ESM.

Line 53: Do you mean details?

Line 54: What do you mean by *"…on pure-resolution approaches…"*?

Line 58: Do you mean downscaling instead of upscaling?

Lines 59-60 *"…Notably, the integration of skip connections within the encoder-decoder architecture, as seen in the YNet model developed by Liu et al. (2020), is a significant advancement…"*: what is meant here? Skip connection was first introduced in U-Net Ronneberger et al. (2015) to prevent the vanishing gradient and information loss through the bottleneck and to transfer fine details from the encoder levels to the corresponding decoder levels. I don't think adding skip connections is a significant advancement now.

Fig.1 (c): It is hard to understand the legend *"meanr testdata obs"*?

Line 106: BCSD is first introduced without an explanation.

Line 124: What type of wind data did you use (speed/direction u/v)?

Line 146: Equations should be numbered sequentially. Please follow the guidelines from https://www.geoscientific-model-development.net/submission.html#math.

I would also add the local predictors such as topographic data to the equation.

Moreover, X and Y are tensors and should be $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{X}$ should also refer to other predictors from ERA5 such as the 10-m wind not just the rainfall. f is a function representing the neural network.

Lines 152-153 *"…This separation into distinct sets for testing and validation enables us to more accurately assess the model's predictive uncertainties across varying data regime…"*:

This allows to assess the model performance for different time periods. If you want to assess the model performance over different regions you should do a spatial split.

Line 157: What is meant by log1p. What does 1 refer to? Is it ln(x+1)?

Line 159: What are the wind vector data (speed or direction) and how do you do the normalization for precipitation? Is it ln(precepetation+1)? What is the reason behind normalizing differently for wind and temperature/humidity? Why were the input variables not normalized to have a zero mean and a standard deviation of one?

Lines 159-160: What kind of humidity and temperature are being used here?

Line 161: Normalization is used to prevent the domination of a subset of the input and to reduce the impact of outliers. In addition, it improves the convergence and keeps the weights in the model within small ranges.

Line 166: I think you mean an encoder and a decoder.

Line 111, subsection 2.1 Data: What is the spatial resolution of the input images? Is it 14×9? and in what coordinate system? Why was an image-based approach chosen over the video-based one?

Lines 168-170 *"…Unlike conventional neural networks that rely on recurrent or convolutional layers, the Transformer architecture is built entirely around attention mechanisms, facilitating direct modelling of dependencies regardless of their distance in the input data…"*: What do you mean by regardless of the distances between the input points? What about the Softmax inside the attention block?

Line 171: What is meant here by processing the entire dataset? I think you mean the entire input data points.

Line 173 *"…This approach not only captures the intricate interdependencies characteristic of climate variables but also introduces flexibility in handling input data of varying sizes…"*: Transformer can't handle varying size due to the positional encoding. Of course, some recent models such as Liu et al. (2022) replaced the relative biases with learnable ones but in principle it can't be stated that transformer can handle a variable input size for spatial data.

Line 176: This should be figure 2 instead of 3.

Line 700, Fig. 2: Do you mean topographic data and positional encoding? Was the first two fully connected layers used without an activation in between? I would also call the decoder a decoder not a resolver to be consistent across the manuscript. In addition, what is the rationale behind the arrangement of the skip connections and layers in a way that is different to the common building block of the transformer which uses: LayerNorm > Attention > LayerNorm > FullyConnectedLayer?

Lines 184-188 *"…Decoder part is designed flexibly that one could implement the desired sub-model for combining the intermediate outputs from the encoder with the topography data and performing a one-step upscaling. As for the downscaling process, the intermediate outputs from the encoder are transitioned to the decoder, which are initially reshaped into* *two-dimensional gridded data before being processed by the decoder…"*: The final output of the encoder not the intermediate output is transferred to the decoder.

Line 186: What is the rationale behind using one-step upscaling instead of the common cascading one i.e., like the standard Upsampling in U-Net.

Lines 198-199 *"…Implementation is carried out within the TensorFlow framework, leveraging its robust capabilities for efficient model training and optimization…"*: I would remove this sentence. Most deep learning frameworks provide efficient training and optimization.

Line198: Did you use Keras API? Please make it clear.

Line 200: I think 60 epochs for an early stopping is a large parameter.

Line 202: Equations should be numbered sequentially. Please follow the guidelines from https://www.geoscientific-model-development.net/submission.html#math.

How is γ defined? There is also a typo, the subscripts *I* and *J* for ground truth should be small letters *i* and *j*.

Please either add batch size B or add a sentence to mention that the batch size is omitted for simplicity.

Line 203: H and W should be in italic H and W.

Line 205 *"…of the training regime…"*: I would write the training procedure/scheme.

Line 207: Do you mean the decoder is trained while the encoder was frozen?

Line 209: Please add a reference to Adam optimizer since there are many versions of Adam. Do you use weight decay parameters?

Line 209 *"…to adjust model parameters effectively during the training process…"*: I would remove this sentence.

Line 211 *"…a choice that significantly enhances computational efficiency…"*: I would remove this part and just mention the time for training.

Lines 212-213 *"…This setup ensures that the model is both accurately and efficiently trained to meet the demands of precise climate data downscaling…"*: Please report the inference time.

Line 257: Where in the appendix?

Line 267: Do you mean in Table 2?

Line 276: Do you mean Fig. 1c?

Lines 274-286: I couldn't find Fig. 1d - Fig. 1f.

Line 308: Where exactly in the appendix?

Figure 1: (a) is missing in the top left corner.

Line 335: Figure 1a does not have observation.

Line 452: I couldn’t find Fig. S1.

Line 464 *“…Echoing the suggestions of numerous studies, adopting reinforcement neural networks, such as generative adversarial networks…”*: Reinforcement learning is not related here. I think you mean generative neural networks such as generative adversarial networks.

Lines 466-467: I would mention diffusion models (Rombach et al. 2022) since they are the state-of-the-art in image generation not GANs anymore.

Line 494: The term DL is first introduced without an explanation. I think you mean Deep Learning.

Line 499: Please add references for U-Net and diffusion models.

Lines 514-517: As I mentioned in the major comments, it is questionable how a model trained on reanalysis data such as ERA5 will perform on climate projections.

Line 518: It is better to publish the preprocessed data and refer to the raw data.

Line 527: Grammar, have.

Line 528: What do you mean by performing the simulation? Did you run simulations?

Line 683: Table 1 Is not mentioned in the text expect in line 267 where the authors meant Table 2. It is also implied from this table that only wind and precipitation were used from ERA5 while temperature and humidity are not mentioned.

Line 685, Table 2: Why are all seasons in bold text except the Summer 1st wet season?

Lines 687-691, Tables 3 and 4: To provide statistically significant results, please run the models with 3 different random seeds and report the mean and standard deviations.

Line 705: figure 3: Do you mean defined in Table 2? It would be helpful to plot the ground truth from TCCIP in this figure.

Line 709, figure 9: I think it is better to put this figure on one page?

References:

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Chiang, C.-H.: EnDeAux_Climate_Downscaling, Zenodo [code], https://doi.org/10.5281/zenodo.10937920, 2024a.

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.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... & Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12009-12019).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł . & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).