Summary of the main contribution

This manuscript proposes a novel deep learning approach to perform precipitation downscaling. Specifically, the authors do this by training five different CNN models (SR-CGAN, Directed/Encoded-Simple, Direct/Encoded-CGAN) to both low- and high- resolution (i.e., 50km and 12km horizontal resolution, respectively) Weather Research and Forecasting (WRF) simulations. The authors apply their methods to a one year WRF simulation and they assess the performance in terms of MSE, probability density function, spatial pattern of some selected summary statistics, and event-based rainfall intensity, duration, size, and total volume. Overall, I found the paper to be well motivated and most of it to be well described. However, I do have some concerns about the evaluation metrics used in this work.

Major points

My major concern has to do with some of the evaluation metrics used in this work:

1. **MSE**: I am not quite sure why the authors define MSE as $\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$, where N is the total number of grids, Y_i is the (fitted) prediction value at grid i, and $\bar{Y}$ is the average prediction across all the grids. A more sensible MSE would be $\frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_{i,\text{Groud Truth}})^2$. Also, I am not sure there is a need to calculate MSE at each timestep unless the authors plan to explore how MSE varies with time. Therefore I would suggest the authors to calculate MSE as $\frac{1}{N\times T}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{i,t} - Y_{i,t,\text{Ground Truth}})^2$, where T is the total time steps during the testing period.

2. **Precipitation Distribution:** I don't think J-S distance used here is very informative, a single number does not tell us *how* two distributions are different (and such assessment can be made, at least qualitatively, using log-pdfs shown in Fig. 6). However, it would be of interest to show spatial maps of J-S distance at each grid cell across different methods as it could potentially provide additional information for sub-region assessment in terms of fitted distributions. I would also suggest to replace the log-pdf curves in Fig. 6 by QQplots, that is, for each region, plot true (empirical) quantiles against the predicted quantiles. Quantile values are more directly interpretable than these log-density curves.

Minor points

- ⋆ page 2, line 42-43 *"Running an RCM is computationally expensive, however, and typically cannot be applied to large ESM ensembles"*:

  Please include reference for the Canadian Regional Climate Model Large Ensemble here, for example, Kirchmeier-Young, M. C., N. P. Gillett, F. W. Zwiers, A. J. Cannon, F. S. Anslow, 2018: Influence of human-induced climate change on British Columbias extreme 2017 fire season. Earth's Future, 7, 2-10. https://doi.org/10.1029/2018EF001050.

- ⋆ page 4, line 100 *"The data used in this study are one-year outputs..."*:

  Did the authors apply their methods to another one-year outputs to check if (qualitatively) similar conclusions can be obtained?

- ⋆ page 13, Table 2:

These information can be well summarized by a scatterplot by putting regions on x-axis and MSEs on y-axis with different color/line symbol combinations for these CNN models.

- ⋆ Figs. 7-9:

  I would suggest the authors to plot relative error maps to better compare between different CNN model fits.

- ⋆ Fig 10:

  I would recommend to use QQplots here for easier comparison.