Review of:
Fast and accurate learned multiresolution dynamical downscaling for precipitation

Recommendation:
Major revisions

Overview:

The authors present a technique for dynamical downscaling using a neural network. They demonstrate that the CNN-based approach can significantly outperform interpolation-based downscaling and CNN-based super resolution (trained on statistically up-scaled data). The CNN-based downscaling approach is many times faster than traditional dynamical downscaling.

Overall, I think this is a significant contribution and am excited to see these CNN-based algorithms come into more widespread use. I do think the paper needs several major revisions. One is related to the MSE evaluation metric. The others generally ask for more exposition about methods and should be fairly easy to address. (Also, regarding methods, thank you to the authors for including a link to their code, this was helpful and contributes to the reproducibility of the work).

Major Comments:

S 2.1: How is scaling/pre-processing/post-processing of the data handled? It looks, based on the code, like the output activation from the CNN is a sigmoid (logistic) function, so how are the precipitation data mapped to/from the (0,1) range? Do you take any steps to account for the very skewed distribution of precipitation? Does the sigmoid output cap the maximum possible output precipitation? If there is a non-linear transform used to map from the CNN outputs back to dimensional precipitation values how does this affect using an MSE loss function? These considerations should be addressed in the manuscript.

Lines 138, 166 and Fig 3: It is not clear how the elevation data are passed to the CNN. Line 138 says it is concatenated to the low-res inputs and Line 166+Fig 3 make it seem that it is only concatenated to features derived from the other fields near the end of the network.

Eqns 1 and 2: Please explain what 'D' is other than just the discriminator output. Your code looks like it is binary cross entropy applied to a sigmoid output with binary class labels. This info should be included here.

Eq 3. Does not seem like an appropriate error metric for this problem. Why is the difference between the SR/downscaling and a spatial mean of precipitation used? Shouldn't this be something like: MSE = (1/N) \sum_{i=1}^N (Y_i^{CNN} – Y_i^{GT})? (GT = ground truth). I believe that is the type of MSE the CNN is optimizing. To me, Eq 3 looks more like the variance of the CNN output computed about the ground truth mean rather than the MSE. Also, just

practically speaking, I don't think adequately estimating total precipitation averaged over all of conus really matters if we can't accurately say where within CONUS it is happening.

S 2.5: No information about the discriminator network architecture is given.

S 3.2: What are the relative numbers of trainable parameters between the models? Do the encoded CNNs improve performance (in terms of the PDFs at least) because of the model architecture difference or simply because they have more convolutional kernels?

Minor Comments:

Title: what is meant by "multi-resolution"? doesn't this only produce 12km outputs?

General: There's inconsistent terminology used in the ML and atmospheric science fields: "up-sampling" and "down-scaling" both refer to resolution increases while "down-sampling" and "up-scaling" both refer to resolution decreases. It would be helpful to make this clear somewhere in the manuscript to prevent confusion for a reader who is only familiar with one of the fields.

Abstract/Line 73: When reading the abstract, it's certainly implied but I don't think it is explicitly stated that a key difference between your method and the CNN-SR method is that one uses high- and low-resolution simulations to train and the other uses a high-res simulation with statistical up-scaling to train.

Line 93: "precipitation images" – is an odd term since the data are really only similar to an image in that they are gridded/raster data. Presumably the CNN is operating on the actual precipitation data and not the RGB images of it.

Line 140: again I don't think referring to the data as images is helpful since they are not images. Something like: "This approach of stacking the variables as different input channels has been used in other downscaling studies…" would be fine.

Line 138: *A friendly suggestion regarding the elevation data*
If you are only concatenating the DEM data near the end of the network like Fig 3 seems to suggest, you may want to consider adding the elevation data earlier in the model. It appears that after the DEM data are concatenated they are only passed through one layer alongside the features learned from the other inputs before the model output. I suspect this could limit the CNN's ability to learn the complicated non-linear relationships between the DEM and the other fields needed to estimate orographic precipitation (letting the information about orography flow through the channel attention blocks might be particularly helpful). I would recommend finding a way to add the DEM data earlier in the network, perhaps you could use convolutional downsampling or a pixel shuffle instead of just 2d-avg to get it down to the same resolution as

the other inputs without losing much information. I think the paper is fine without doing this of course, since you have already achieved very good performance.

S 2.3: Did you provide any source of random variability to the CGAN?

Table 2: Units?

SS2.5 It feels disorganized to introduce the GAN model in a section labeled "loss functions." I suggest just having a section early on that briefly introduces each model to help the reader keep everything straight. Then proceed to discuss specifics of implementation / loss functions etc.

Very Minor / Typos:

Line 10: Awkward sentence

Line 21: change to something like: "ESMs cannot *fully* resolve cloud processes." the way it's currently worded makes it sound like the models don't include clouds at all

Line 45: one number is formatted with a comma and one without

Line 57: GAN-based SR doesn't necessarily improve pixel accuracy over a conventional CNN, they improve feature loss or realism. Maybe just be clear about what you mean by "accuracy"

109: "convections"→"convection"

F1: Thanks for including this diagram, very helpful

L 194-195: I don't think this sentence is correct

L 204: "less good" → "worse"

L 225: wow that's fast! I'm excited to see algorithms like this used operationally