This is an interesting and useful study examining the comparative ability of several neural network (NN) architectures to improve short-term/fine-scale precipitation forecasts from NWP models. The paper demonstrates measurable improvements relative to the baseline while also highlighting clear limitations in the ability of any of the architectures to correctly predict more intense rainfall. I believe this paper is suitable for publication with the following suggested minor revisions.

Line 35: Please expand the abbreviation 'GRU'

Line 81–84: It seems unlikely to me that all 143 forecast variables contribute significantly to the rain prediction. I wonder whether training speed and robustness — including reducing the risk of overfitting — could be significantly improved by substantially reducing the variable set. I don't expect the authors to perform new runs for this paper, but some comments on the question of "how many variables is too many?" would be welcome.

Section 2.3: It would be helpful to know what fraction of the roughly 20,000 observations had any rain at all. An overall histogram or CDF of rain rate for the entire observational data set could also be useful. If the fraction of rain is low, then I would imagine that the NNs are effectively being trained to *not* predict rain by default, which might help explain the low biases? Would it be desirable to bias the training sample toward raining cases to reduce this tendency?

Line 135: If I'm reading this correctly, all models were trained for a fixed number of epochs without early stopping and without specific tests for convergence. Is this correct? Please consider clarifying the training procedure with a few additional details.

Line 137: Was an optimal learning rate determined experimentally? Is it to be expected that the same learning rate is optimal for several very different NN architectures?

Section 3.3.3: Heading should be "Equitable Threa*t* Score"

Table 2: Finding the highest median score in each row, I see that Deconv3L did best for 0.2 mm/hr, Deconv1L did best for 0.5 mm/hr, and U-Net did best for 1 mm/hr. And all three did better than CGAN for these rain rate thresholds. Yet CGAN did best in MAE and LEPS scores. The lack of a consistent pattern of any one method standing out by most or all metrics seems to me to be a significant finding, but I don't think this comes through clearly as clearly as it could in the discussion or in the conclusions.

Line 236: Should "overcast" be "overforecast"?

Line 272: "Considering the excellent performance shown by the Deconv3L model …" I suggest "superior" (in the sense of "better by some metrics than the other models") rather than "excellent", as the latter word implies that there's only limited room for improvement. The MAE score of 0.5 strikes me as good in relative terms but not entirely satisfying in absolute terms.

Lines 334-335: To "future research", I suggest including an evaluation of which of the 143 forecast variables are actually relevant to the models' performance—e.g., by withholding one at time and seeing how performance changes. As mentioned above, I suspect that excluding irrelevant features from the data set—and thereby improving the signal-to-noise ratio at the inputs—might improve the robustness of the training that is possible with a limited data set.