DeepPrecip: A deep neural network for precipitation retrievals.

King et al. (2022) submitted to AMT

Review by:

Anonymous Reviewer

Introduction and Recommendation:

Precipitation is a vital measurement for the earth sciences and humanity because of its direct impacts on human life and property though flooding as well as its in direct effects of water resources for human consumption and agriculture. Despite the importance of precipitation measurement, measuring precipitation on a global scale has been a challenging endeavor and has been continuously pursued by NASA for more than two decades now (i.e., TRMM & GPM). The most direct and likely most accurate method of measuring precipitation is through rain gauges, where the amount of liquid water equivalent precipitation can be measured. The drawback of gauges is that they are only point measurements and precipitation has high spatio-temporal variability. Thus, remote sensing efforts can alleviate in-situ measurement drawbacks, but can suffer from their own suite of issues (e.g., calibration, DSD assumptions etc.). Thus there is plenty of room for improvement of these remote sensing techniques.

The authors contribute to the remote sensing literature by presenting a remote sensing method to derive the surface precipitation accumulation. To do the retrieval, the authors they take advantage of a quickly growing method in meteorology/atmospheric sciences, machine learning (c.f., Figure 1 Chase et al. 2022). More Specifically, the authors use vertically pointing radars (i.e., MRR) paired with surface rain gauge measurements from various global locations to train a deep learning retrieval to map measured radar values to the surface precipitation value. In their analysis they show that their new method, DeepPrecip, is able to outperform a random forest method and more simple power-law relationships (which are predominately the legacy method of going from radar measured data to precipitation rate/accumulation).

The paper is generally well written, and the authors are knowledgeable on the methods used in the paper. Furthermore, this paper does fit the scope of AMT and would be a valuable contribution to the literature after addressing the few comments I have made below. I am formally designating these comments as **major** revisions because I am not sure if they will take more than 2 weeks to implement or not (this is the designation between major and minor in my head).

Major comments:

Units of the accumulation data

The main concern I have is with the main result of the paper and the units of the data in Figure 4. To me these accumulation values seem unreasonably small, which makes me concerned there is some sort of error in either a unit conversion or these are truly just really light precipitation events. Let me explain my reasoning.

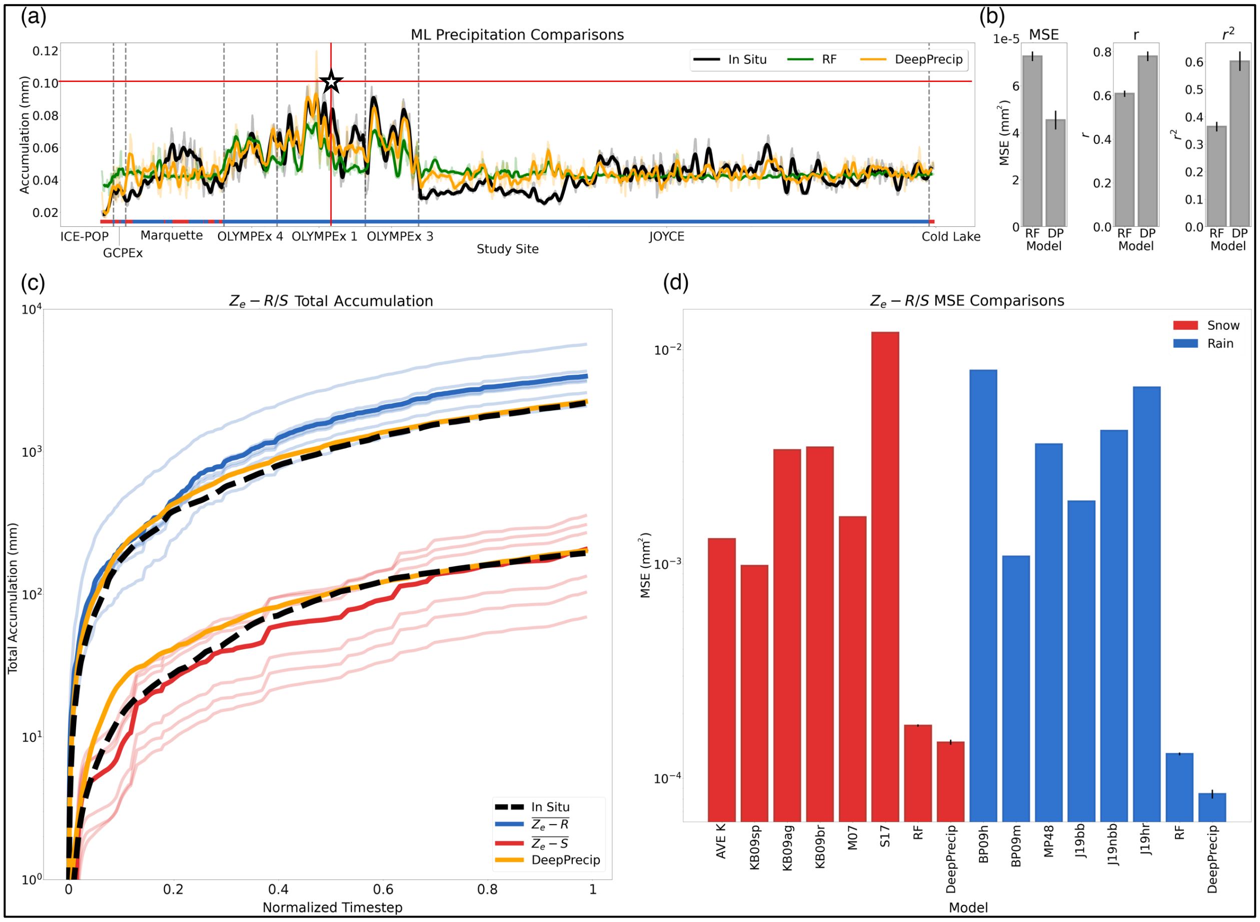

Figure 1: Figure 4 from the paper with my annotation of the maximum accumulation I read from the figure (Black Star and lines).

The maximum accumulation reported in Figure 4a (and Figure 6a) shows a 20 min accumulation of 0.1 mm. I am more familiar with English units, so 0.1 mm is 0.004 inches of precipitation. 0.004 inches of precipitation is less than the precision of a common ASOS tipping bucket (1 tip in a tipping bucket gauge is 0.01 inches of precipitation). Thus most of the precipitation events shown in Figure 4 are instances where the amount of precipitation in 20 mins is less than 1 tip of an ASOS rain gauge. I hope you see my concern if these accumulations are correct. If the maximum precipitation amount is less than 1 tip of a tipping bucket gauge how representative is the data the authors present compared to the 'global' distribution of precipitation, especially convective precipitation.

While we are discussing Figure 4a, do you have any thoughts of why for JOYCE and Marquette the ML predictions are basically anchored around 0.04 mm?

One last comment on Figure 4, (specifically c). I am confused how this plot was made. What is a normalized time step? Could you help readers by explaining a bit more how this plot was made in the text?

Transparency of which dataset results are from:
While I am happy the authors made sure they spent the time to explain their hyperparameter search and some details of their training/test splits, there is no explicit comment of which dataset is being shown in the results section. This is vital to any machine learning paper. The authors must state if the results being shown are from the training or the test dataset. This will allow readers to assess if the results are an unbiased assessment of skill or if they seem to be overfit.

This gets more challenging since the authors did a k-fold cross validation approach, I am unsure which fold they used to show the results. Please explain.

Table 2 and Literature power-laws:

Upon inspection of the references provided for Table 2, I noticed that some of these are not specifically K-band relationships. For example Kulie and Bennartz (2009) derive relationships for W-, Ka- and Ku- but not K. Similarly, Matrosov (2007) is for Ka and W. Lastly, Marshall and Palmer (1948) is a Rayleigh power law. While it might seem like using a Ka-band relationship for K-band is harmless, issues arise when non-Rayleigh conditions are encountered (e.g., particle sizes are similar to the wavelength), which tend to coincide with large precipitation rates. You should acknowledge that this could be a source of error in your analysis and might be an unfair comparison for your discussion on lines 195 – 209 (Figure 4cd).

Lastly, you state in the table caption and the discussion on line 112 that all the relationships are K-band, which is incorrect. To prevent future readers from inaccurately using the reported relationships in your paper on their K-band radars, please correct this mistake.

Random Forest model details:

The authors mention a random forest model that was based on previous work, but no citation is provided for this model. Given that this random forest model is involved in the primary conclusions of this paper, there is more detail needed. The current description on Lines 146 – 149 is insufficient for reproducibility do not include any of the details of how big the random forest is. Please provide the citation where this model was developed. If there is no citation, please provide more specific details on the random forest model. Also, please note if this model was re-trained on your current data or is it still using the X-band snowfall relationships from GCPEX.

Minor comments:

Scaling of data:
There was no discussion on if you scaled the features of the ML model. It is common practice to scale data to have mean 0 and variance 1 in machine learning so that the ML model doesn't unintentionally use a variable with a larger absolute value. For example the dynamic range of radar reflectivity is -10 – 40 dBZ. While the range of temperature is 233 – 313, and the range of doppler velocity is -5 – 5 m/s. See how these three all vary on a different order of magnitude?

Did you end up scaling your data? Or did you use batchnorm in training? (I did not see this a parameter in Table 5). Please comment on this.

Citation issues:

It would seem there was an issue with LaTex building the document, all of the parenthetical citations do not correctly put citations into parentheses. This made reading some parts of the paper more difficult. Please be sure to use \citep[e.g.,][]{Paper} to correctly get the formatting to work. (or \citet for inline citations).

Line by Line Comments:

Please note that word change suggestions are suggestions! Please do not feel pressured to accept my recommendations.

Line 16: This is an example of the citation issue noted in the minor comments.

Lines 60: Be careful here. In my head liquid water content is usually the water content per cubic meter (e.g., g/m^3). Might be good to use a different word here, "records precipitation accumulation" something like that.

Line 84: could you spell out what TMP and WVL are? This is the first time they are defined

Figure 2: I assume darker colors mean higher density? You might want to either include a colorbar somewhere or write it in the caption. Could you note in the caption that the wind velocity is vertical wind velocity and which direction negative is? (is negative wind velocities up or down?). This confused me at first because I thought it might be the horizontal wind velocity, but then I didn't know how to interpret negative values. Why is the unit in m/s on Figure 2 for wind velocity, but in Pa/s in Table 3?

Lines 93 – 94: How much data was not used because of the 5 m/s wind threshold. It is my experience that some of the strongest precipitation events occur coincidently with strong winds. You might want to comment how this effects the total scope of precipitation events you are training your model on.

Lines 95 – 102: If you were to extend this work in the future, it might be good to use wet-bulb temperature as a way to split when it is raining vs snowing (Sims and Liu 2015).

Line 152: 90/10 split is sufficient usually, but could you comment on how using a non-shuffled dataset could have seasonality issues? What I mean by that is that often times field campaigns are centered on the event they wish to capture. Thus the bookend times (near the beginning of a campaign and near the end of the campaign), precipitation might be reduced (coming into or out of a 'dry' season). This could be a problem if all of your test splits have weak precipitation events.

Line 276: What p-value and statistical test was used to make the significant conclusion? Please refrain from using the word significant unless you used a statistical test to determine significance.

Lines 306-307: What do you mean by 'assimilate non-attenuated near surface radar data" in the context of spaceborne radars? As you noted before the blind-zone is an issue because of clutter, not attenuation. I am a bit confused by this statement.

Lines 315-316: Just because an echo is > 3 km does not mean it is convective. There are plenty of GPM and CloudSat profiles that have stratiform echoes reaching all the way up to the tropopause (~10 km in the mid-latitudes). Also, the planetary boundary layer in most locations is likely not extending up to 3km. I would guess maybe 1-2 km on average. But I am not an expert in boundary layers. Be careful in the statements here.

Data Availability: Why not cite the NASA data websites here? E.g., https://ghrc.nsstc.nasa.gov/uso/ds_details/collections/gpmgcpxC.html this would be helpful for people to grab the data.

References:

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., & McGovern, A. (2022). A Machine Learning Tutorial for Operational Meteorology, Part I: Traditional Machine Learning, *Weather and Forecasting* (published online ahead of print 2022). Retrieved Jul 18, 2022, from https://journals.ametsoc.org/view/journals/wefo/aop/WAF-D-22-0070.1/WAF-D-22-0070.1.xml

Sims, E. M., & Liu, G. (2015). A Parameterization of the Probability of Snow–Rain Transition, *Journal of Hydrometeorology*, *16*(4), 1466-1477. Retrieved Jul 18, 2022, from https://journals.ametsoc.org/view/journals/hydr/16/4/jhm-d-14-0211_1.xml