GPROF-NN: A neural network based implementation of the Goddard Profiling Algorithm

Response to reviewer comments

1 Comments from reviewer 1

The paper presents a convolutional neural network architecture for IR precipitation retrieval over Brazil. The training data are from IR and GPM combined retrievals. The framework is extended such that it can provide uncertainty of the retrievals. The estimates are compared with ground-based gauge data to validate the retrievals. The paper is well written and is of high-quality. I have the following comments.

1.1 Major comments

Reviewer comment 1

Validation only with a month of gauge data is not sufficient for clamming those improved results in the abstract. Seasonal to annual validation results are needed to make those claims.

Author response:

We agree with the reviewer that a more thorough evaluation of the retrieval accuracy over extended periods is desirable. However, we also want to point out that the evaluation presented in Sec. 3.1 in the original version of the manuscript covers the full year of 2020. Thus, the evaluation does already cover longer time scales than the month used to evaluate the precipitation accumulations. Nonetheless, it is true that an analysis of the accuracy across different time scales is missing from the manuscript.
One difficulty with extending the evaluation against gauge measurements is the storage capacity required to store input data and results. For example, input and output data of the Hydronn retrievals for one month require 2.5 TB of storage.
We therefore propose the following extension of our evaluation scheme, which will allow assessing the retrieval performance across seasonal time scales within the constraints of the compute resources that are currently available to us:

1. We will extend the evaluation against the GPM combined measurements to cover the full year of 2020. We will compare our retrievals to GPROF and HYDRO. We choose GPROF instead of IMERG for this comparison because the retrievals can be directly collocated in time with the reference data, which is not possible for the gridded IMERG data. The GPROF retrievals therefore constitute a stronger baseline for instantaneous precipitation estimates. We choose not to include PERSIANN CCS because the data is only available at hourly resolution and comparison against the instantaneous reference measurements would make the product look overly bad.

2. We will extend the evaluation against the gauge measurements to also cover June 2020.

This extended evaluation scheme allows us to show the robustness of the accuracy of our retrievals for both instantaneous and accumulated measurements. Since the intended application of algorithms are near real-time retrievals, we consider the assessment of the retrieval accuracy across annual time scales outside the scope of this manuscript. We will also extend the discussion of the retrieval accuracy to reflect those points.

Reviewer comment 2

A single storm retrieval is missing. It is imperative to show the output of the algorithm in retrieval of a single or multiple storms and compare the results with the combined GPM retrievals as a reference. One retrieval snapshot speaks very clearly about the skill of the algorithm in reconstructing the training data and retrieve spatial structure of precipitation.

Author response:

We will add retrieval results for an overpass of the GPM core observatory over a mesoscale convective system. We also add a comparison of the the retrieval results to GPM PMW retrievals and the HYDRO algorithm. To further illustrate the capabilities of our retrieval, we will also include a video of the retrieval results at 10-minute resolution over 24 hours as a supplement with the manuscript.

Reviewer comment 3

Error metrics are only represented cumulatively. The expectation is that paper presents the quality of retrievals for an individual storm in terms of detection accuracy (e.g., probability of detection, miss) and then focuses on estimation quality metrics at different time scales from a storm scale to monthly and seasonal.

Author response:

As stated in response to reviewer comments 1 and 2, we will extend the evaluation of instantaneous precipitation estimates and include an assessment of the retrieval accuracy for a single storm case.
Estimation quality metrics are already reported for hourly, daily, and monthly time scales in Tab. 3. The extension of the evaluation scheme proposed in response to comment 1 will also allow to assess the retrieval accuracy across seasonal time scales.

Reviewer comment 4

This needs to be clarified whether the training data were only over Brazil or not. If this is the case, then the provided improved statics are not of surprise. This issue needs to be stated in the abstract.

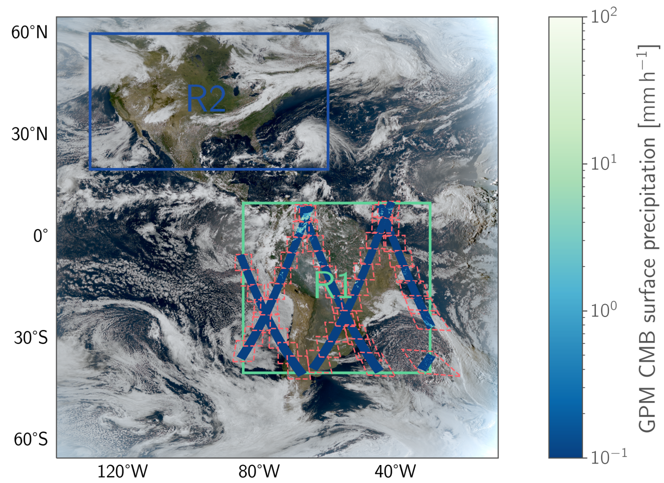

Figure 1.1: GOES-16 true-color composite from September 23, 2019 (generated using the `natural_color` composite in satpy (Raspaud et al., 2021)). The rectangle R1 marks the domain over South America, which was used for the extraction of training and testing collocations between the ABI on GOES 16 and GPM CMB. Dashed polygons show the boundaries of the training scenes extracted for this day together with the collocated GPM CMB results. The rectangle R2 marks the secondary domain which is used as an additional test domain to assess the impact of the spatially limited training domain.

Author response:

We will add a statement to the abstract stating that the training data is restricted to South America and a figure showing the region used to extract the training data for the retrieval.

Moreover, we think the reviewer's suggestion that our reported improvements are 'not of surprise' brings up an interesting question. Namely, whether these improvements are due to the more representative training data or the more expressive statistical models used by Hydronn. To investigate this question, we will add a further evaluation of the retrieval accuracy over a separate region (R2 in Fig. 1.1) over the northern hemisphere.

Reviewer comment 5

The way the paper explains the Bayesian retrieval is confusing. First, what is the prior distribution? Just obtaining uncertainty of estimates does not mean that the approach is Bayesian, and we can call the distribution a posterior. We can quantify uncertainty in a frequentists sense. It seems that the approach counts the number of retrievals associated with Tbs within bins. Then the bin with maximum is labeled. The problem is then defined as classification problem and the output of the softmax function is considered as the posterior distribution of the retrievals. Even though, I found the approach creative, I am not convinced that it is a Bayesian approach.

Author response:

The connection between probabilistic neural network retrievals and Bayesian retrieval methods has been shown in Pfreundschuh et al. (2018). A reference to this article is included in l. 135 of the first version of the manuscript, which also state that the distribution of the training data in this case corresponds to the a priori distribution.
It is of course possible to interpret the probabilistic results in a frequentist sense, however, the Bayesian framework is, at least in our experience, more common for inverse problems in satellite remote sensing. It also has the advantage that it highlights the dependence of the retrieval results on the a priori assumptions, i.e., the training data of the neural network.
Since the relation between training data and a priori distribution of Bayesian retrievals is fundamental to our work, we will revise the manuscript to better convey the significance of the training data in the Bayesian retrieval framework.

Reviewer comment 6

It is claimed that spatially aware CNNs provide more accurate retrievals than pixel-level DNNs. The reason is not discussed, and no evidence is provided.

Author response:

The evidence for the higher accuracy of CNN retrievals stems from a preliminary study to which a reference is provided in the manuscript. Since it seem that this has not been made sufficiently clear, we will rewrite the section to more clearly state where these results can be found.

Reviewer comment 7

In equation 2, when the prior probability approaches to a small number, the likelihood ratio can be extremely large. The correction numbers in Fig. 4 are too large. Please explain why such a large difference might exist in the retrievals that need such a large correction factor. For correcting probability distribution we can use a simple CDF matching!

Author response:

Upon revisiting the likelihood ratios, we have come to the conclusion that the calculation presented in the first version of the manuscript was not correct. Instead of using the training data to calculate the correction factors, we will recalculate the probability ratios using a priori distributions derived from retrieval results. This will likely decrease the magnitude of probability ratios. However, large probability ratios are still possible whenever the a priori distribution of the retrieval approaches zero. Therefore, differences in the measurement characteristics between the GPM combined retrieval and the gauge measurements can still lead to large probability ratios.

This is certainly a drawback of our approach. However, the CDF matching approach proposed by the reviewer is typically used to correct scalar retrieval results. We are, therefore, not aware of a way to apply the method the probabilistic output provided by our retrievals.

Reviewer comment 8

The resolution of IR is higher than microwave data. In this sense, you have redundant samples. How were those samples treated in the training?

Author response:

We did not treat these samples in any particular way. Since all training samples are revisited multiple times during the training anyways, the induced redundancy is unlikely to be an issue for the retrieval.

Reviewer comment 9

Explanation of the uncertainty quantification is too complex. Please consider simplifying the text and provide improved explanations.

Author response:

We will rewrite the section describing the approach to quantify uncertainties aiming to make it easier to understand.

1.2 Minor comments

Reviewer comment 1

Why both the second and third configurations are needed. They are just different in resolution. Line 160. Provide reasoning.

Author response:

We will elaborate on the motivation for the third configuration retrieval configuration.

Reviewer comment 2

Line 185. The range is too wide! The training GPM combined precipitation can range from 0.1 to 200 mm/hr. Why 1000 mm/hr?

Author response:

The range that we employ for the probability bins is certainly excessively wide. While it is true that this range could be reduced, this is very unlikely to affect the performance of the retrievals in any way. We thus don't consider this to be an issue.

Bibliography

Pfreundschuh, S., Eriksson, P., Duncan, D., Rydberg, B., Håkansson, N., and Thoss, A.: A neural network approach to estimating a posteriori distributions of Bayesian retrieval problems, Atmos. Meas. Tech., 11, 4627–4643, https://doi.org/10.5194/amt-11-4627-2018, 2018.

Raspaud, M., Hoese, D., Lahtinen, P., Finkensieper, S., Holl, G., Proud, S., Dybbroe, A., Meraner, A., Feltz, J., Zhang, X., Joro, S., Roberts, W., Ørum Rasmussen, L., strandgren, BENR0, Méndez, J. H. B., Zhu, Y., Daruwala, R., Jasmin, T., mherbertson, Kliche, C., Barnie, T., Sigurðsson, E., R.K.Garcia, Leppelt, T., TT, ColinDuff, Egede, U., LTMeyer, and Itkin, M.: pytroll/satpy: Version 0.33.1, https://doi.org/10.5281/zenodo.5789830, 2021.