Review of: "Scalable radar-driven approach with compact gradient-boosting models for gap filling in high-resolution precipitation measurements" by Peter Lünenschloß et al.

Overview

In this paper, a machine-learning driven method for filling data gaps in rain gauge networks using radar data is introduced. The method is computationally efficient and is designed to prevent calibration leakage. It is evaluated by comparing radar-derived data with real rain gauge measurements.

The paper is written well and contains a good overview of the overall scientific problem and the existing solutions for it. The new method introduced in this work is clearly described. In my opinion, however, the analysis and discussion of results is not as thorough. The output of the new method is evaluated in terms of several statistical parameters, but these are not compared to alternative methods (though many are mentioned in this work). The physical and application-related implications of the accuracy of the method are also only briefly discussed. It would be good if the authors could improve on these aspects before publishing this work.

General/major comments

1. The authors took great care to put their methods in broader context and provided a thorough overview of existing solutions for measurement gap filling problems. In the light of this, I find it surprising that the results are not explicitly compared to any other method: authors just provide the values of several standard statistical tests and claim that they are good. I fully understand that implementing alternative approaches for the particular example that the authors used here might be laborious, but perhaps some key quantitative results from some of the numerous cited works could be quoted in the main text? The claim that the values of the statistical parameters used to evaluate the new method's performance are "good", must, after all, be based on a comparison of some kind? Even if some of the statistical tests are standard, basic comparisons to other works would save a non-expert reader from having to dig into all the references.

2. Figures 4 and 7 give an impression that the rainfall from some short and intense rain periods (which authors attribute to "highly convective or erratic rainfall events") can be significantly underrepresented. This may be a problem for some of the applications, but has not been explicitly quantified. Is the performance seen in Figure 4 and Figure 7 typical? Is there an overall low bias in rainfall predictions? Is there a bias in case of short periods of intense rain? These questions seem particularly relevant, since authors suggest to use the new method for runoff simulations and flash-flood prediction.

3. The authors claim that the method has low computational costs and good scalability. Unless I have missed something, the only concrete data regarding overall computational cost was the statement that "end-to-end processing for one station" required around 15 min on a single CPU core "with 30 GB of assigned RAM". This information is sufficient to convince me that computational costs are indeed quite low, and thus the method is in principle suitable for processing large data sets and/or real time applications. However, more information would be needed if claims about scalability are made. Firstly, was the model really constrained to a single CPU core, or was it simply not (explicitly) parallelized? Modern programming languages often have significant implicit parallelization capabilities. Secondly, while 30 GB of memory was assigned, how much was actually used? This is important, because if the model indeed runs on a single CPU core and requires 30 GB of memory, it is a rather memory-intensive application. Both modern workstations and HPC systems typically have way less than 30 GB of memory *per CPU core*, and would, in this case, struggle to process many stations in parallel. Finally, since not all

CPU cores are equally powerful, it would be good to specify the model of CPU used here.

Minor/specific comments

1. Table 2: RMSE, at least according to the standard definition, has the same dimension as the measured quantity. However, no units are given for RMSE values in the table. Is RMSE given in mm here? This should be specified clearly.

2. Figure 5: I do not think that the names of the variables are sufficiently self-explanatory here. If they are standard, maybe an appropriate reference could be provided here? Otherwise consider introducing them more explicitly (in an Appendix, perhaps?).

3. Figure 6 caption: Perhaps *rows* refer to increasing gap lengths? Since columns are labelled (a) - (c) and seemingly refer to something else? Also, I think the gaps lengths should be explicitly specified.

4. Figure 7: The title mentions 95% prediction interval, and the caption talks about 90% confidence interval. Which one is actually represented in the figure? Also, it looks like all the information in Figure 4 is also given in Figure 7. If that is indeed the case, is Figure 4 really necessary? If it just there for easier reference in the corresponding part of the main text, maybe different stations or time intervals could be chosen for the two figures, then more information could be shown without increasing clutter and complexity?

Minor typos and suggestions

1. L17: "rely on high frequency rainfall to capture". Perhaps "rely on high frequency rainfall measurements to capture"?

2. L122 and elsewhere: in my opinion, the usage of the words "exemplar" and "exemplary" is somewhat odd throughout this paper. In English, the primary meaning of the word "exemplary" is "very good", "an example others should follow", using it simply to say "used as an example" is a lot less common. It is even more unusual to use the word "exemplar" where "example" would fit.

3. L150: I suggest to replace "chosen such that" with "chosen so that".

4. Table 2 caption: The words "german weather service" are not capitalized here, but they are everywhere else. Also, the word "Partner" is capitalized with no apparent reason.

5. L411: Duplicated sentence.