Review on “A climatological benchmark for operational radar rainfall bias reduction” by Ruben Imhoff et al.

This paper presents a novel approach for deriving an Adjustment Factor (AF) map, which is able to reduce the biases that inevitably affect any Quantitative Precipitation Estimation (QPE) algorithm that is solely based on weather radar observations. Personally, I find particularly interesting the following idea of the approach: each pixel of the (spatially varying) AF map, (10-y climatologically) varies as a function of the day of the year. The Adjustment Factors are, in fact, based on a 10-year data set from various operational ground-based networks: (~330) daily and (~30) automatic rain gages and 2 weather radars. However, real- time application of the method is problematic, as it is acknowledged by the authors themselves. Hence, I agree with the authors, the proposed method could become a benchmark for radar meteorology in flat terrain: this is the reason why I suggest to add one or two evaluation Sections in Chapter 3.
The paper has high value from a practical point of view, especially for hydrological, meteorological and climatological services: I think they will appreciate the pragmatic approach, like I did. It discusses a relevant topic, it is clearly structured, comprehensive and well written; most of it could be really published as is. Can the proposed method become a better benchmark than the current MFB adjustment run operationally? To answer this relevant question, the authors should add one (or two) Section(s) dealing with daily (and/or hourly) comparison of the networks of rain gages and weather radar. That explains my MAJOR REQUEST (see page 3 and 4) to please insert two additional “Evaluation Sections” based on daily (all CARROTS gages, ~360) and hourly (only 31 automatic gages) sums, which will complement the (less relevant, I dare saying) information of the (current) Sec. 3.2 “Annual rainfall sums”

I also have a few comments/suggestions (see below), which together with some (possible) additional literature, could link this interesting study and benchmark in a flat terrain to the challenge of using weather radar also in mountainous regions.

I congratulate the authors for the interesting and pleasant study; I strongly invite them to add the suggested new Sections “Daily and Hourly QPE performance” (or something similar) and the corresponding relevant information in the revised manuscript.

Locarno Monti, 9.4.2021

Yours faithfully,
Marco Gabella

COMMENTS/SUGGESTIONS
I start with figures, since I have enjoyed them so much! For a long time, I have not review a manuscript with figures with such high quality. I have enjoyed the color palette and the nice paper by Crameri et al. 2020, which I was not aware of. I again congratulate the authors and thank them for the great care they have devoted to their interesting and informative figures.

Line 98 and figure 2:
nice figure! Please complement it with a Table showing not only the 5 (DJF MAM JJA SON) slope values (which are now in the bottom right of the figure and you can put in the Table) but also

- the 5 Pearson correl. Coefficients,
- the 5 sample size values (Number of samples,)
- the 5 values of $1 / F_{MFB}$ as defined in eq. 2

Figure 4:
nice figure, especially 4a). I appreciate the Log-transformation of the Bias Factor in the ordinate axis of 4a). Similarly, I highly recommend to use a Log scale for the y axis of 4c)

Important: is fig. 4c consistent with 4a? The minimum Bias correction Factor in 4a (“country-wide) takes place somewhere in June. Why does the minimum Bias correction Factor in 4c correspond to April?!? And May is smaller than June, too?? What am I missing? Sorry, I am lost here.

Line 27: this schematic introduction, with a subdivision of the sources of error in three classes remind me of a contribution presented at the IV International Symposium on Hydrologic Applications of Weather Radar, San Diego, 1998**, with selected papers subsequently published on a special issue of JGR 2000, same issue of Borga et al., 2000 in your references. Well the paper I am referring to** is also listed again below, line 43 (being aware that I am biased, you see, Locarno-Monti, CH, it is where I have started learning radar meteorology back in the 90’s …)

Line 33: a fully automatic and operational correction based on a mesobeta vertical reflectivity profile has been successfully running in Switzerland for almost 20 years now! Hence, I propose to add Germann and Joss, 2002, to the list

Line 43:
Regarding range-adjustment and AF map in not-too-complex terrain, one of the oldest contribution that I have in mind is the one by Koistinen and Puhakka, AMS radar Conf. 1981, which can be complemented by Koistinen et al., AMS radar Conf. 1999 and Michelson and Koistinen, ERAD 2000(I had the chance to attend both of them :-) In those years Joss’ idea of two additional predictors in complex terrain has also been presented (Gabella, Joss and Perona, JGR 2000 **special issue): min. height of radar visibility and terrain altitude. The Adjustment Factors were derived by means of a non-linear Weighted Multiple Regression (WMR). In the US, I remember Seo et al. 2000, J. Hydrometeorol. In Gabella 2004 IEEE, the WMR is trained during the 1st day of the event and then applied during the following days ...

Sec. 2.2.2
This is the HEART of your paper! I mean, the spatio-temporal variability of of F_{clim} whith respect to the simplistic F_{mfb}: you have chosen 31-day and 10 year. So, please anticipate that you will discuss further the corresponding implications in Sec. 3.4 and Fig. 7 (I was a bit worry when reading the first time … until I have reached such Sec. 3.4 …)

Line 144:
Please say something more about KGE and/or show its formula. Is 1 the optimal value? What does minus infinity mean? [Fig. 6(i)]. By the way: does its formulation with limitations explain in Fig. 6(k) why the biased raw radar product (KGE=0.84) perform better than F_{mfb}?

Line 170: see line 132, it would be nice to refer to Germann and Joss, 2002.

Sec. 3.4 Fig. 7
Please elaborate further and discuss the variability of F_{clim} whith respect to the simplistic F_{mfb}: what if you used a “seasonal” window (91-day running average) for smoothing instead of the current monthly one (31-day)? What would you suggest to a national meteorological service with a “short” 5-year archive? If you had a 30-year (or 20-year) archive, would you still use 31-day? Would you try to see what happens with a 7-day running average?

Line 247: yes, in San Diego 1998**, it was shown that in mountainous terrain, the Adjustment Factor map have a dependency on the height of visibility from the radar; (not only Borga, Anagnostou and Frank, JGR 2000, but also Gabella, Joss and Perona, I dare offering)

Lines 265-270: Indeed! If you were interested to see the performance of the “best-on-average” QPE product, you could read Panziera et al., 2018 (Int. J. of Climatology). Evaluation there is based on “leave-one-out” approach (see also col. 4 my Table, next page).

Somewhere in the Conclusions: the new Dutch radar are dual-pol, are not they? Could you be interested in a Random Forest approach? (see Wolfensberger et al., 2021).

MAJOR REQUEST:
I find Sec. 3.2, which deals just with annual amounts, not adequate to charcterize the radar-gage comparison. I highly recommend the authors to analyze all wet days (hours) in their large and precious data set and provide some scores. For instance, a score could be simply the daily (hourly) Root Mean Square "Error" (I call "error" the simple difference between Ra and Rc, Ra and Rmfb, Ra and Ru) divided by the conditional mean daily (hourly) precipitation rate (mean of Ra, considering only wet hours, see column 2 in the exemplificative table below). Such normalized and dimensionless score is often called Fractional Standard Error.
We know that verification of precipitation data is not straightforward. We know that the radar-gage comparison is an intriguing task, especially as far as the interpretation is concerned. We are aware of the fact that we do not know the truth; all measurements are subject to errors. However, to simply omit the comparison and avoid the (sometime difficult) interpretation is not the solution, I think. Rather, I think that new Sections with daily (and/or) hourly radar-gage comparison would make this manuscript richer and more interesting.
As stated, I propose something very simple: the (inter-)annual (variability of the) daily/hourly FSE for the three products when compared with the reference at the ground, Ra, which is the gage amounts. Fig. 3 and 4 clearly suggest a 4-season stratification when preparing this kind of Tables. Maybe the authors do not want to insert in the papers all the eight Tables, just describe similiarities/differences among them. Or maybe they will just do the exercise for daily OR hourly amounts? (4 Tables). Or just winter versus summer season? However, I am sure that at least one or two of such tables will provide interesting information and give a better overview of the MFB versus CARROTS performance.

For the sake of comparison, I have prepared a similar Table for the summer season in the complex orography region of the Swiss territory. The 3rd column shows the FSE for our solely radar QPE product. The 4th column, the FSE of the radar-gage merging product (CombiPrecip, see Sideris et al. 2014), derived by excluding the gage at hand, when deriving the CombiPrecip value for the pixel that contains the gage. The last column shows the (ultra-optimistic) FSE value that one would derive by using simply the CombiPrecip maps and neglecting the obvious fact that the pixels that contains the gage is highly influenced by the gage value itself

Year	Summer average hourly Rain Rate Conditional upon G≥0.1 mm/h	"Radar" Fractional Standard Error (FSE)	"CombiPrecip Leave-one-out" FSE	"CombiPrecip" FSE
2005	1.339 mm/h	0.63	0.42	0.32
2006	1.301 mm/h	0.85	0.62	0.55
2007	1.514 mm/h	0.57	0.44	0.30
2008	1.329 mm/h	0.71	0.49	0.46
2009	1.539 mm/h	0.57	0.43	0.34
2010	1.363 mm/h	0.64	0.45	0.37
2013	1.387 mm/h	0.44	0.33	0.23
2014	1.306 mm/h	0.51	0.37	0.24
2015	1.292 mm/h	0.63	0.41	0.36
2016	1.415 mm/h	0.47	0.34	0.24
2017	1.535 mm/h	0.43	0.29	0.21
2018	1.571 mm/h	0.47	0.32	0.26
2019	1.731 mm/h	0.39	0.27	0.21
2020	1.478 mm/h	0.44	0.30	0.21

The first six lines refer to the old, Doppler, single-pol. network of 3 radars. In this period only 70 telemetered rain gauges were available in the whole Country.
The other eight lines refer to the renewed dual-pol. network: 3 radars in 2013, 4 radars since 2014 and 5 radars since 2016. The number of telemetered rain gauges has increased considerably starting from 2012 to reach to remarkable total of 266 in 2016.

As stated, to obtain the RMSE in mm/h it is enough to multiply the FSE values in columns 3, 4 and 5 by the normalization values listed in the 2nd columns.

Having said that, a straightforward (and somehow trivial) interpretation is the following:
Better radar hardware, an increased number of radars together with an increased number of gages improves QPE performance.

Other considerations regarding radar-only QPE:
- for the old network, best [worst] performance has occurred with the strongest [weakest] conditional average rain rate: FSE=0.57 <-> E{G}=1.5 mm/h; FSE=0.85 <-> E{G}=1.3 mm/h
- for the new network, best [worst] performance has occurred with the strongest [weakest] cond. average rain rate, too: FSE=0.39 <-> E{G}=1.7 mm/h; FSE=0.63 <-> E{G}=1.3 mm/h

Something similar (but not identical for the period 2006-2010), can be observed in the leave-one-out (or the optimistic, last column) evaluation of the radar-gage merging product CombiPrecip.

It would be nice if you could derive FSE values also in a leave-one-out mode. Maybe it is too much work for you and not convenient from a cost/benefit viewpoint?

Note that MSE is the sum of the square of the Mean Error plus its Variance. Hence, RMSE can be heavily affected by the BIAS component. There is a score which is perfectly orthogonal to the BIAS in dB. It is called “Scatter”, it is also expressed in dB, it is a weighted average of the Log-transformed Cumulative Distribution Error Function. Unfortunately, it cannot deal with zeros. It has been presented at ERAD2004, see page . If you were interested, we are willing to share Python, IDL, Matlab, R, routines for it (probably not for this paper, rather for future evaluation?)

CITED REFERENCES

J. Koistinen, and T. Puhakka, “An improved spatial gauge-radar adjustment technique,” Preprints of the 20th Conf on Radar Meteorology, Boston, MA, pp. 179-186, 1981.

J. Koistinen, R. King, E. Saltikoff, and A. Harju, “Monitoring and assessment of systematic measurement errors in the NORDRAD network,” Preprints of the 29th Conf. on Radar Meteorology, Montreal, Canada, pp. 765-768, 1999.

M. Gabella, J. Joss, and G. Perona, “Optimizing quantitative precipitation estimates using a non-coherent and a coherent radar operating on the same area,” *J. Geophys. Res.,* vol. 105, pp. 2237-2245, 2000.

D. B. Michelson, and J. Koistinen, “Gauge-radar network adjustment for the Baltic Sea experiment,” Phys. Chem. Earth (B), vol. 25, pp. 915-920, 2000.

D. J. Seo, J. P. Breidenbach, R. Fulton, D. Miller, and T. O’Banon, “Real time adjustment of range dependent biases in WSR-88D rainfall estimates due to non-uniform vertical profile of refractivity,” Journal of Hydrometeorology, vol. 1, pp. 222-224, 2000.

U. Germann; Joss, J. Mesobeta profiles to extrapolate radar precipitation measurements above the Alps to the ground level. J. Appl. Meteorol., 41, pp. 542–557, 2002.

M. Gabella M, 2004: Improving Operational Measurement of Precipitation Using Radar in Mountainous Terrain - Part II: Verification and Applications, IEEE Geoscience and Remote Sensing Letters, **1**, 84-89.

L. Panziera, Gabella M., Germann U., Martius O., 2018: A 12-year radar-based climatology of daily and sub-daily extreme precipitation over the Swiss Alps, Int. J. Climatol., 7-21.

D. Wolfensberger, Gabella M., Boscacci M., Germann U., Berne A., 2021: Rain Forest: A random forest algorithm for QPE over Switzerland https://amt.copernicus.org/preprints/amt-2020-284/