Author response to referee #4 (Marco Gabella)

HESS-2021-105

Ruben Imhoff

Ruben.Imhoff@deltares.nl

May 21, 2021

Dear reviewer,

We would like to thank you for your interest in our work and the enthusiastic reaction to our manuscript. In addition, we would like to thank you for sharing some results for Switzerland. That has given us a good example of what the extra analysis in the manuscript could look like.

With four constructive and elaborate reviews, we think we are very well served by our reviewers. Your comments have been valuable and have helped us to improve the manuscript.

Below, we give a response to the given suggestions. We have placed the reviewer's comments in black font and below that, our response in blue font for clarity.

Sincerely,

Ruben Imhoff, Claudia Brauer, Klaas-Jan van Heeringen, Hidde Leijnse, Aart Overeem, Albrecht Weerts and Remko Uijlenhoet

General comments

I find Sec. 3.2, which deals just with annual amounts, not adequate to characterize 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?)

We would like to thank the reviewer for this suggestion and the detailed description of a possible extra analysis. On top of that, it is great to see the example for Switzerland. Thanks for sharing this! We have decided to implement an analysis taking into account the FSE per QPE product (unadjusted, MFB-adjusted and corrected with CARROTS) for every rainy hour in the 10-year dataset. Besides that, we have also decided to keep the analysis per year as it was already present in the original manuscript. This decision was based on the comments of the other reviewers and after some adjustments (see our responses to the other reviewers), we think it is worth keeping from a more hydrological perspective (as this clearly shows the volume differences over a longer period).

We have applied the reviewer’s suggestion for all seasons in the ten-year dataset. This is based on a grid cell-based comparison between the reference and the QPE products, which is subsequently averaged over the land surface area of the Netherlands to provide one FSE value. Note that we have decided not to go for a leave-one-out mode or an analysis of both hourly and daily rainfall intensities. Instead, we have tried to keep it straightforward with only an hourly intensity approach as described below. This will change some text and it adds a table. We propose to place the table in the appendix as additional information (we will show the winter and summer seasons, as this fits together in one table) and to place a figure with a scatter between the daily rainfall sums of the reference and the QPE products (similar to Fig. 2) in the text, because this directly visualizes the information in the FSE table. First of all, we propose to add a paragraph to Sec. 2.3:

“**2.3 Hydrometeorological application**
Both bias adjustment methods were applied to the ten years (2009–2018) of R_U. In order to provide a hydro-meteorological testbed, both the CARROTS and MFB adjusted QPE products (from here-on referred to as R_C and R_{MFB}, respectively) were validated against the reference rainfall. First, this was done on country level. The estimated daily rainfall sums for all grid cells within the land surface area of the Netherlands were compared to the reference in a similar way as the comparison in Fig. 2. To subdivide these results per year and season, an additional hourly rainfall sum validation took place as well. The results of this analysis can be found in the appendix and the analysis took place as follows: for every rainy hour (when the sum of at least one grid cell > 0.0 mm), we compute the Root Mean Square Error (RMSE) by squaring the differences between the three QPE products (R_U, R_C and R_{MFB}) on the one hand and the reference on the other, and taking the average of these squared differences over all grid cells within the land surface area of the Netherlands. Subsequently, the RMSE was averaged over all rainy hours in that season and year. Finally, the seasonal mean RMSE was divided by the average hourly rainfall rate for that season and year, resulting in the Fractional Standard Error (FSE) score. The FSE score was calculated for every season in the ten years to be able to compare the seasonal performance of the hourly rainfall estimates of R_C to R_U and R_{MFB}.

Second, the annual rainfall sums for the twelve basins in Fig. 1 were compared with the reference. In addition, R_C and R_{MFB} were used as input for the rainfall-runoff models of twelve basins (a combination of catchments and polders) in the Netherlands (Fig. 1). […]”

With the new results, Sec. 3.2 in the results will change to:

"3.2 Evaluation of the rainfall sums

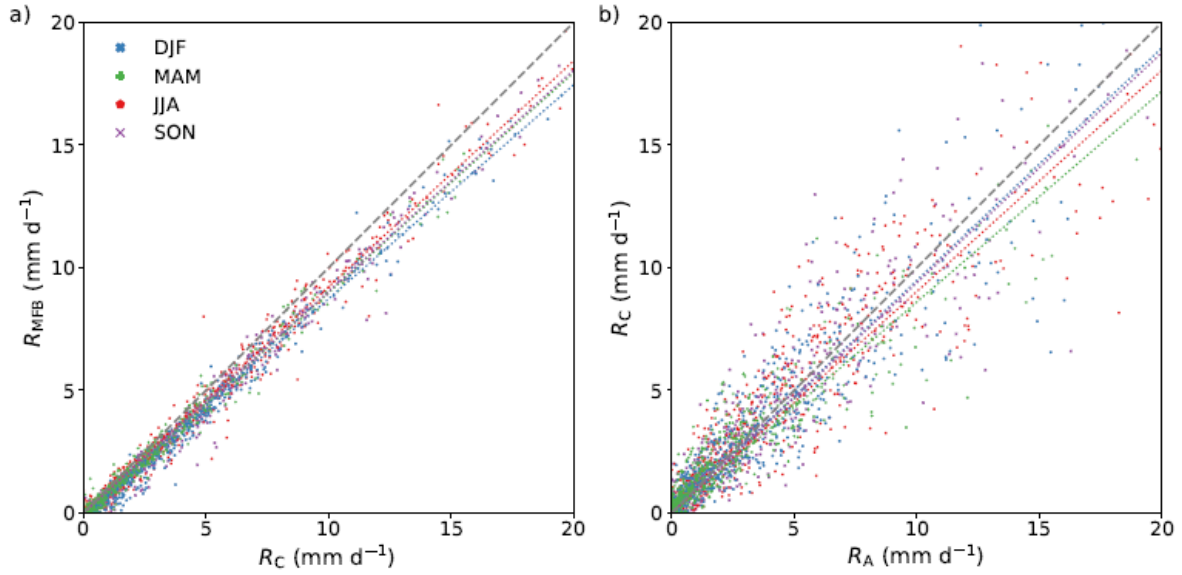

Figure 5. Comparison between the reference rainfall (R_A) and the two adjusted radar QPE products: (a) R_{MFB} and (b) R_C). Shown are the daily country-average rainfall sums based on ten years (2009–2018), classified per season. The slope, Pearson correlation and sample size per season are indicated in Tab. 2.

Table 2. Statistics of Fig. 5. Indicated are the sample size, the Pearson correlation and the slope of a linear fit between the reference and the two adjusted radar QPE products (R_{MFB} and R_C). This is indicated per season and for all seasons together (Total).

		Slope		Pearson correlation	
Season	Sample size	R_{MFB}	R_C	R_{MFB}	R_C
DJF	902	0.87	0.95	0.99	0.92
MAM	920	0.90	0.86	0.99	0.92
JJA	920	0.92	0.90	0.99	0.91
SON	910	0.90	0.94	0.99	0.93
Total	3652	0.90	0.92	0.99	0.92

The MFB adjusted QPE (R_{MFB}) significantly reduces the systematic bias of R_U (Fig. 2), from a 55% underestimation on average for the Netherlands to 10% (Fig. 5a and Tab. 2). The remaining bias in R_{MFB} results, however, generally from a systematic underestimation of the reference rainfall. The overall underestimation is less for R_C (8%, Fig. 5b), but results from estimation errors that are either under- or overestimating the reference rainfall. The spread in Fig. 5b is significantly higher than in Fig. 5a, indicating that the country-wide QPE error of R_C is often higher than for R_{MFB}. The yearly FSE in Tab. 3 clearly indicates this too, with a systematically higher FSE for R_C than for R_{MFB}.

An advantage of the MFB adjustment is that it corrects for the circumstances during that specific day and thus also for instances with overestimations (Fig. 4a). On a country-wide level, this is clearly advantageous, also compared to CARROTS (Fig. 5). The negative effect of the spatial uniformity of the factor, however, becomes apparent in Fig. 6, which compares the annual precipitation sums of the two adjusted radar rainfall products with the reference and R_U for the twelve basins. For all basins, both adjusted products manage to significantly increase the QPE towards the reference. However, for nine out of twelve basins, R_C outperforms R_{MFB} (Fig. 6e). Exceptions are Beemster, Luntersebeek and Dwarsdiep, where the performance of both products is not that different.

Differences between the performance of R_C and R_{MFB} become most apparent for catchments that are located more to the edges of the radar domain. For instance, R_{MFB} for the Aa and Regge catchments, which are located in the far south and east of the country, still underestimates the annual reference rainfall sums with on average 20% for the Aa (MAE = 151 mm and mean annual R_A = 761 mm) and 13% for the Regge (MAE = 103 mm and mean annual R_A = 776 mm), while this is on average only 5% (both under- and overestimations occur) for R_C (Fig. 6b and c).

Table 3. Country-average Fractional Standard Error (FSE) between the hourly reference rainfall (R_A) and the three QPE products (R_U, R_{MFB} and R_C) per year for the winter and summer seasons. The FSE was only calculated for hours where the country-average rainfall rate was larger than 0.0 mm h^{-1}.

			FSE		
Seaon	Year	Avg P rate (mm h^{-1})	R_U	R_{MFB}	R_C
DJF	2009	0.32	1.10	0.49	0.74
	2010	0.26	1.23	0.61	0.82
	2011	0.38	1.12	0.50	0.73
	2012	0.36	1.09	0.51	0.65
	2013	0.30	1.04	0.56	0.90
	2014	0.33	1.06	0.51	0.72
	2015	0.34	1.04	0.51	0.84
	2016	0.34	1.15	0.61	0.84
	2017	0.37	0.56	0.32	0.44
	2018	0.37	1.22	0.65	0.76
JJA	2009	0.33	1.18	0.80	1.08
	2010	0.43	1.34	0.71	1.02
	2011	0.37	1.31	0.78	1.03
	2012	0.36	1.19	0.72	0.99
	2013	0.36	1.34	0.86	1.20
	2014	0.33	1.37	0.91	1.28
	2015	0.44	1.24	0.69	1.08
	2016	0.30	1.46	1.00	1.46
	2017	0.37	1.29	0.76	1.09
	2018	0.34	1.26	0.78	1.20

The MFB-adjusted QPE performs better for the Beemster polder, Dwarsdiep polder (Fig. 6d) and Luntersebeek catchment (Fig. 6a) due to their location in the radar mosaic. The Luntersebeek catchment (central Netherlands, Fig. 1) is located closer to both radars. There, R_{MFB} generally performs better and sometimes even overestimates the true rainfall, which is consistent with Holleman (2007). The performance of R_{MFB} for the Dwarsdiep catchment is similar to its performance for the Linde catchment (both in the North of the country), but R_C shows more variability in the error from year to year for the Dwarsdiep catchment (Fig. 6d), leading to a better relative performance of R_{MFB}. The CARROTS QPE tends to overestimate the rainfall amount of the three aforementioned basins (Beemster, Dwarsdiep and Luntersebeek) for some years (e.g. by 16% for the Luntersebeek in 2016). Overall, the performance of R_C and R_{MFB} are not that different for these three basins, with on average just a lower MAE for R_{MFB} than for R_C for the polders Beemster and Dwarsdiep (Fig. 6e).

Summarizing, the CARROTS factors have a clear annual cycle, with generally higher adjustment factors further away from the radars (Sec. 3.1). On average for the Netherlands, the MFB adjusted QPE outperforms the CARROTS-corrected QPE. However, the spatial variability in the CARROTS factors, in contrast to the uniform MFB adjustment, results in estimated annual rainfall sums for the

twelve hydrological basins that are generally closer to the reference (for nine out of twelve basins) than with the MFB-adjusted QPE, especially for the east and south of the country. This effect is expected to become more pronounced when the adjusted QPE products are used for discharge simulations."

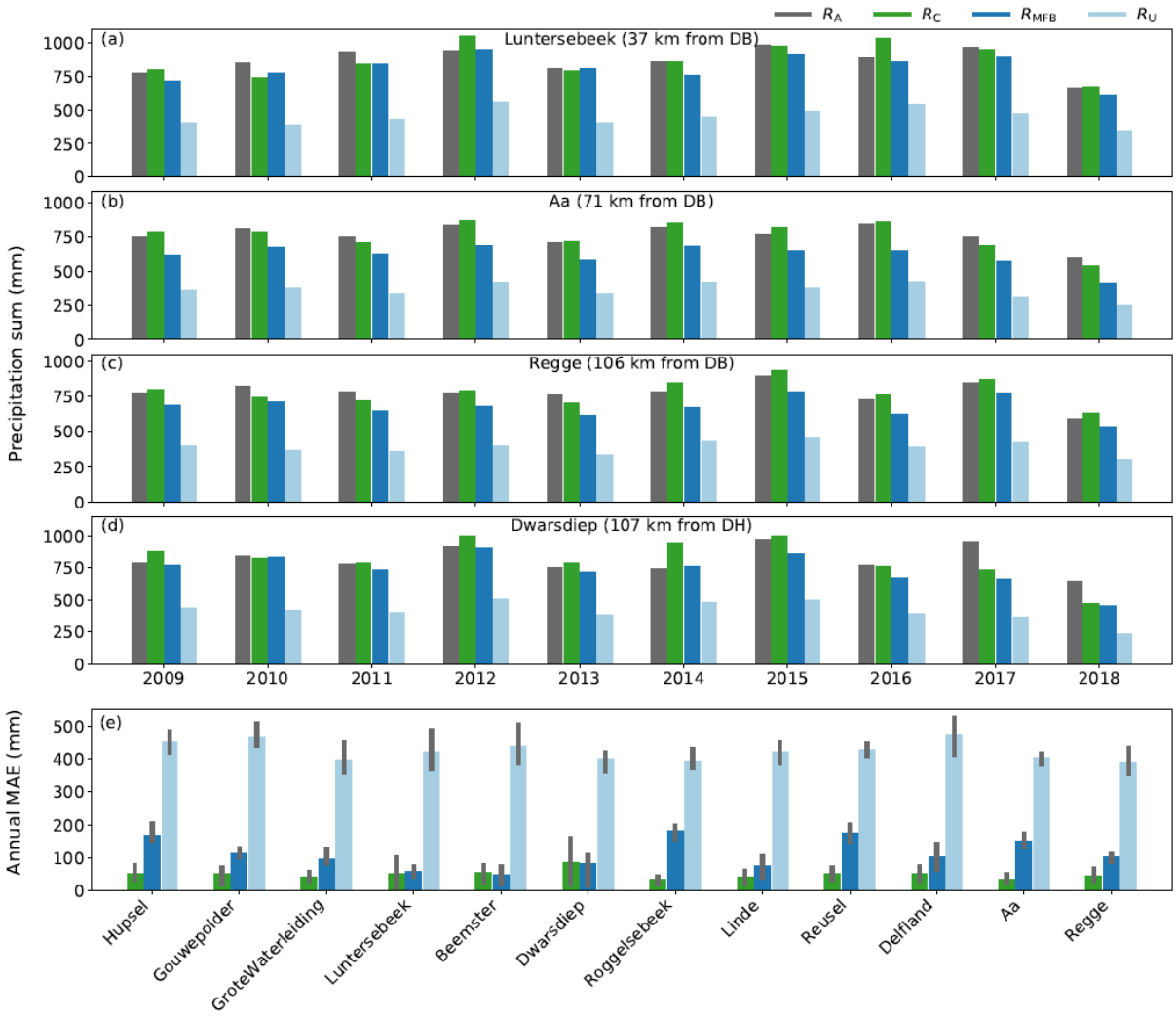

Figure 6. Effect of the adjustment factors on the catchment-averaged annual rainfall sums. (a – d) The results for a sample of four catchments that are spread over the country (and thus the radar domain): (a) Luntersebeek, (b) Aa, (c) Regge and (d) Dwarsdiep. Shown are R_A (grey), the estimated rainfall sum after correction with the CARROTS factors (R_C; green), the estimated rainfall sum after correction with the MFB adjustment factors (R_{MFB}; dark blue) and the rainfall sum with the unadjusted radar rainfall estimates (R_U; light blue). The distance between the catchment center and the closest radar in the domain is given in the title of subfigures a-d (DH is Den Helder and DB is De Bilt). The radar in Herwijnen, which replaced the radar in De Bilt in January 2017, is not included here, because this radar was operational for the shortest time in this analysis. (e) the mean absolute error of the annual precipitation sum between the QPE products and the reference rainfall sum (R_A). The vertical grey lines, per bar, indicate the IQR of the MAE based on the ten years.

Finally, the results from the proposed changes (notably Fig. 5 and Tab. 2) should also be summarized in the conclusions. We propose two changes: (1) In lines 285 – 287, we will add that we have also looked at daily and sub-daily rainfall estimates for the land surface area of the Netherlands. (2) In

the QPE performance paragraph (lines 292 – 296) we will add: “On average for the Netherlands, the MFB adjusted QPE outperforms the CARROTS-corrected QPE.”

Specific comment and technical corrections

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 / FMFB as defined in eq. 2

We would like to thank the reviewer for this suggestion. We have added the information to a table and the figure and table will look as follows:

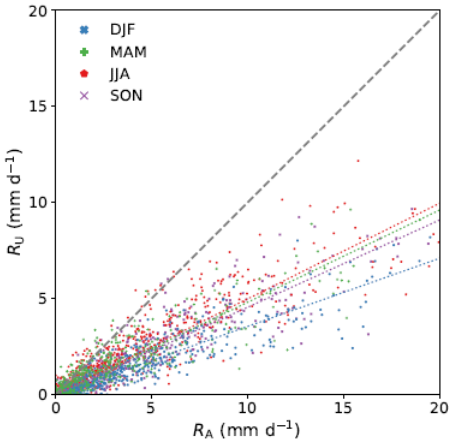

Figure 2. Scatter plot indicating the systematic discrepancy between the reference rainfall (R_A) and the unadjusted radar QPE (R_U). Shown are the daily country-average rainfall sums based on ten years (2009–2018), classified per season. The slope, Pearson correlation and sample size per season are indicated in Tab. 1.

Table 1. Statistics of Fig. 2. Indicated are the sample size, the slope of a linear fit between the two rainfall products (R_A and R_U) for all observations and the Pearson correlation. This is indicated per season and for all seasons together (Total).

Season	Sample size	Slope	Pearson correlation
DJF	902	0.35	0.90
MAM	920	0.48	0.89
JJA	920	0.50	0.89
SON	910	0.45	0.92
Total	3652	0.45	0.89

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.

It is a bit hard to see due to the small range (1.5 – 3.0), but the y-axis of Fig. 4c is also on a logarithmic scale. Regarding the lower factor in April than in June (well spotted!), Fig. 4c shows the effective adjustment factor for the superimposed grid cell of KNMI station De Bilt in the middle of the country. The reason for the difference with 4a is twofold: (1) the effective factor is, as described in the figure caption: "based on only the rainfall sums within that month, the effective adjustment factor for that month, which roughly coincides with the factor for the 15th of the month in the CARROTS method." Hence, the April factor roughly coincides with the 15th of April in Fig. 4a and the June factor with the 15th of June in 4a. From June to July, the factor increases and therefore this 'average' factor for June is somewhat higher than the minimum at the end of May / start of June in Fig. 4a. This should be a minor detail, but indicates why June is not lower than April. (2) At station De Bilt, and its superimposed grid cell, the minimum takes place earlier, somewhere between April and May. This location is most comparable to the location of catchment Luntersebeek (the yellow – orange line in Fig. 4a) which shows a similar behavior. Thus, this grid cell is not entirely representative of the behavior that we see in many other catchments (and grid cells further away from the radars).

We propose to add a sentence to the caption, describing this: "(c) Dependency of the monthly adjustment factor on the estimated 0^{o}C isotherm level for KNMI station De Bilt and the superimposed grid cell of this station. Depending on the location in the radar composite, the minimum CARROTS factor can take place in a different month, but is always between April and June. Note that for this analysis, the adjustment factor was based on only the rainfall sums within that month, the effective adjustment factor for that month, which roughly coincides with the factor for the 15th of the month in the CARROTS method. The grey bars indicate the interquartile range (IQR) for that month, based on the spread in hourly 0^{o}C isotherm level estimates (the horizontal bars) and the sensitivity to leaving out individuals years in the ten-year period for the factor derivation (vertical bars)."

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 ...

As response to the three suggestions above: Good suggestions, we will add these references!

Sec. 2.2.2: This is the HEART of your paper! I mean, the spatio-temporal variability of of Fclim whith respect to the simplistic Fmfb: 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 …)

Another reviewer also pointed this out. In response to that reviewer, we indicate in which sections and figures the results of Sec. 2.4 can be found. In addition, we propose to refer to Sec. 2.4 in Sec. 2.2 to indicate the presence of a sensitivity analysis.

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 Fmfb?

This was also mentioned by multiple reviewers. Again, thanks for pointing this out. In our attempt to keep the text as brief as possible, we have overlooked the need to introduce the KGE metric more elaborately. We plan to elaborate the sentence with: "The resulting discharge simulations were validated for the same period and 5-min timestep using the Kling-Gupta Efficiency (KGE) metric (Gupta et al., 2009):"

"$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$,

$\alpha = \frac{\sigma_s}{\sigma_o}$,

$\beta = \frac{\mu_s}{\mu_o}$,

with r the linear correlation between observed and simulated discharge, α the flow variability error between observed and simulated discharge and β the bias between mean simulated (μ_s) and mean observed (μ_o) discharge. σ_s and σ_o are the standard deviation in the simulated and observed discharge. The KGE metric ranges from -∞ to 1.0, with 1.0 being a perfect agreement between observations and simulations. In this study, the discharge simulated with R_A as input was regarded as the observation."

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

Thanks for the suggestion, we will do so.

Sec. 3.4 Fig. 7
Please elaborate further and discuss the variability of Fclim whith respect to the simplistic Fmfb: 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?

Starting with the bad news, we cannot comment on the minimum archive length or running average window size for different archive (of national meteorological services), because we have not tested this. Also in line with the comments of reviewer #1, we propose to add a few sentences to lines 248 – 251 in the discussion section:

"That CARROTS is relatively insensitive to such minor changes in the composite or the year-to-year variability of rainfall, is likely a result of the ten-year archive that has been used. The sensitivity analysis in Sec. 3.4 has shown that leaving individual years out of the archive hardly influences the CARROTS factors. Nevertheless, based on the current analysis we cannot conclude what the minimum number of years in the archive has to be to obtain stable CARROTS factors that are similar

to the factors derived in this study. This is a recommendation for future research as such a sensitivity analysis can point out what the minimum archive length has to be to calculate the CARROTS factors."

Regarding elaborating on the moving window size analysis, we propose to extent that paragraph a bit to:
"The use of a different moving window size hardly influences the CARROTS factors for moving window sizes of two weeks or longer, but this does not hold for moving window sizes of a day or, to a lesser extent, one week (Fig. 7a). The factor derived with a moving window size of one day fluctuates heavily from day to day. This suggests that the adjustment factor is still quite sensitive to individual events in the 10-year period, when a moving window size of seven days or less is used. Moving window sizes of more than a month (6 weeks and 2 months were tested here), lead to similar CARROTS factors as with a 1-month (31-day) moving window size, but somewhat more smoothed. A similar effect likely takes place for a seasonal (3-month) moving window. For larger moving window sizes (half a year to a year, for instance), we expect that the seasonality in the factor is lost and that an average correction factor remains.

In contrast to this, the differences between these six sets of CARROTS factors (Fig. 7a) lead to minimal variations in the simulated discharges for the Aa catchment, when these factors are used to adjust the input QPE (Fig. 7b). Differences in timing and magnitude (0.2–0.3 mm d^{-1}) are visible during peaks and recessions, for instance in early April. However, these are small compared to the differences between the model runs with R_C and R_{MFB} (Fig. 6). However, the use of a window size of 1 day or of, to a lesser extent, a week clearly leads to more fluctuations in the CARROTS factor (Fig. 7a) and can therefore influence the rainfall estimation for individual events. For quickly responding catchments and urban catchments, this could still lead to different results. Concluding, a 31-day smoothing of the climatological adjustment factor is warranted."

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)

Good suggestion, we will add it.

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).

Thanks, that is indeed an interesting paper! In addition, reviewer #1 asked us to elaborate a bit on the results in these lines. We propose to change the paragraph to:

"As mentioned in the previous paragraph, the climatological adjustment factor is not calculated for the current meteorological conditions and resulting QPE errors, which could lead to considerable errors during extreme events. Nonetheless, this is also the case for the MFB adjustment technique (Schleiss et al., 2020). The absolute errors for the 10 highest daily sums in this study for the Aa and Hupsel Brook catchments (one of the largest and the smallest catchment in the study) are similar for the MFB and climatological adjustment methods, with on average a 20% difference with the reference (this would have been 50–60 % without corrections). In most of these events, both R_C and R_{MFB} underestimated the true rainfall amount. However, for a small number of these top 10 events, the QPE products overestimated the true rainfall amount. This occurred more frequently with CARROTS (25% of the cases) than with the MFB adjustment (15% of the cases). Note that for individual events in these twenty extremes, the errors can still reach 48% for the QPE adjusted with CARROTS and 64% for the MFB adjusted QPE.".

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).

That would be very interesting to test in the Netherlands! We saw that the method tends to overestimate low intensity precipitation (typical Dutch winter precipitation), but perhaps if the model is trained on the Dutch data, it turns out differently. Although this is outside the scope of this work, it is definitely worth trying it at some point.

References

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Schleiss, M., Olsson, J., Berg, P., Niemi, T., Kokkonen, T., Thorndahl, S., Nielsen, R., Ellerbæk Nielsen, J., Bozhinova, D., and Pulkkinen, S.: The accuracy of weather radar in heavy rain: a comparative study for Denmark, the Netherlands, Finland and Sweden, Hydrology and Earth System Sciences, 24, 3157–3188, https://doi.org/10.5194/hess-24-3157-2020, 2020.