Author response to all referees

HESS-2021-105

Ruben Imhoff

Ruben.Imhoff@deltares.nl

June 11, 2021

Dear reviewers,

We would like to thank you for your interest in our work and the enthusiastic reactions to our manuscript. 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 ordered per reviewer. We have placed the reviewer's comments in black font and below that, our response in blue font for clarity. In a separate PDF, we have attached the revised manuscript including track changes.

Sincerely,

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

Author response to anonymous referee #1

General comments

1) Do you think, MFB could be improved by dividing the NL into spatial segments (like the classical moving window) or even depending on the distance to the radar, or is the density of automatic stations to low? (The argument that MFB is limited to one factor for a whole country does not hold in general.)

   Thanks for suggesting this. We have tried this before and it generally gave better results than with the country-wide MFB-adjustment factor. The question then remains how large these regions should be. With smaller regions, the local MFB factor is better able to correct for the local spatial (and distance-dependent) errors. That possibility of doing that depends, as the reviewer also indicated, on the local density of automatic gauges. Knowing that it frequently occurs that one or multiple gauges give(s) no or unreliable values, the regions should not become too small in order to have sufficient gauges within the region. If the regions contain a low number of rain gauges, there will be the risk that none of the gauges catch rainfall (especially during convective events), and thus no adjustment factor will be derived. With a country-wide MFB adjustment factor, this probability is lower, but that comes at the price of not at all taking into account the spatial errors in the radar QPE.

   We think that this option is no matter what worth mentioning, so we propose to briefly mention it in the discussion section, where we will add it to the lines 243 – 245, as: "MFB adjustment of radar rainfall fields is still the most frequently applied adjustment method (Holleman, 2007; Harrison et al., 2009; Thorndahl et al., 2014; Goudenhoofdt and Delobbe, 2016). The results indicate that this choice may be reconsidered, at least for the Netherlands and in case a country-wide or large-region adjustment factor is applied. More regionalized MFB adjustments are possible, but depend on the density and availability of the automatic gauge stations."

2) Does the RA adjustment eliminate spatial dependencies of the resulting QPE? To be more precise: Is the quality of RA depending on the distance to the radar site? And if so - would CARROTS allow for the derivation of an improved adjustment procedure?

   Given the dense manual rain gauge network, 1 gauge per 100 km^2, used in the spatial adjustment on a daily basis, we expect spatial dependencies of the resulting QPE in R_A are relatively small, especially on a daily basis. In earlier work (Overeem et al., 2009a), we noticed underestimation of extreme rainfall for short durations, which can be attributed largely to remaining errors in radar data. The combination of a daily spatial adjustment and an hourly mean-field bias adjustment is probably not sufficient to remove these errors completely on sub-daily timescales. Rain-induced attenuation typically results in underestimation of heavy rainfall at longer range from the radar, but not systematically at the same locations. Changes in the vertical profile of reflectivity, which can result in systematically lower rainfall estimates at further range from the radar, are likely easier to adjust for using rain gauge data. It is typically associated with longer lasting stratiform events for which the daily spatial adjustment factor is more suitable compared to more localized short-duration convective events. Overall, the quality of R_A could definitely be improved for sub-daily rainfall, but large systematic differences in space are not expected.

3) What's the effect of temporally present spokes (positive and negative) in the historical data set?

That is very minor, mostly due to the 10-yr mean and the moving window of a month. We hope that the presented sensitivity analysis gives an indication of the stability of the factor. Nevertheless, we have also tested the effect of including 2008 in the factor derivation. This year gave an odd result in the KNMI products, as also described in Sec. 2.1: "The year 2008 is actually the first year in the KNMI archive of both data sets, but it was left out of the analysis here. R_U for this year showed a significantly different behaviour than the other years, especially during the first half year in which the product rarely underestimated and frequently even overestimated the rainfall sums. The reason for this behaviour is not yet fully understood. KNMI (2009) reported that spring was exceptionally dry in the north of the country and that the months January and May were among the warmest on record. On some days with overestimations, clear bright band effects were visible in the radar mosaic, which may have contributed to the systematic differences." This significantly different first half year does impact the factor derivation and resulted in somewhat lower correction factors for the first six months. Although the effect is not a major one, we could observe it in the results, while similar effects are not present for e.g. the (extreme) dry year 2018.

4) I propose to add a figure showing the pixel-based differences between MFB and RA as well as RC and RA on a 5-min-basis, e.g. box-whisker (NL mean or catchments) and map with median and percentiles. This would help understanding the effects on discharges in different regions.

Thanks for this suggestion. We do agree that the QPE validation can be more elaborate. Instead of the reviewer's suggestion, we propose to make two different changes: (1) a change to Fig. 5 including the absolute error with the reference for all catchments. See below (our response to line 178) for the renewed figure and corresponding text in the results. (2) A scatter plot, similar to Fig. 2, showing the performance of both R_{MFB} and R_C, and a table showing the Fractions Standard Error (FSE) based on the hourly rainfall sums of all QPE products and the reference for the land surface area of the Netherlands. We will show this per year and season to give an indication of the (standardized) seasonal and annual variability in the rainfall estimation error. We refer to our response to reviewer #4 for the specifics.

5) What is the performance in heavy rain situations? Despite mean numbers, please comment on the effects.

We agree that we can say a bit more about heavy rain situations. We propose to change "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). 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." into:

"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 based on the reference 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 to 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.".

6) Could a dbz-dependent factor improve CARROTS?

That is an interesting idea! We can imagine that especially for extreme (summer) rainfall events, a dBZ-dependent factor can significantly improve CARROTS, because in these situations the QPE error is generally higher (Schleiss et al., 2020), while the CARROTS factor is based on the average error for that time of the year, which is partly (mostly even) based on less extreme events. The remaining question is then whether the 10-year dataset contains enough (relatively) extreme events for a stable and reliable derivation of a dBZ-dependent factor for high intensity rainfall. There will be sufficient lower-intensity stratiform events, but we expect that the current seasonality in the factor already gives a reasonably good correction for the errors for such (winter) events.

We propose to add a sentence at the end of the paragraph at lines 265 – 271: "A way to better correct for biases during extreme events could be to derive either different Z-R relationships, depending on the type of rainfall, or dBZ-dependent correction factors, which could be derived in a similar way to the CARROTS derivation method. Whether this works or not for extreme events depends on the number of such events in the available historical dataset."

Specific comments and technical corrections

Lines 62-63: I would expect the need to adjust the analysis BEFORE spatially dislocate the reflectivities to avoid adjustment with climate correction factors that are not specific to the original measurement location. Please clarify, why a post-processing should be preferable. I don't see any advantage in that.

Thanks for this remark. This statement was based on the previous paragraph where we stated: "when the adjustment method changes the spatial structure of the original radar rainfall fields (kriging and Bayesian methods), this may impact the continuity of the rainfall fields over time and thereby also the radar rainfall nowcasts (Ochoa-Rodriguez et al., 2013; Na and Yoo, 2018)." However, for MFB adjustments and CARROTS you are absolutely right. Hence, we propose to change "(2) is available in real time so that it can be used operationally for postprocessing of radar-based rainfall forecasts, such as nowcasting" into "(2) is available in real time so that it can be used operationally for radar-based rainfall forecasts, such as nowcasting".

Line 87 - "distance-weighted interpolation": Please comment a bit more detailed.

We agree that we can elaborate a bit more on how this method was applied. The original description can be found in Overeem et al. (2009b). We have tried to concisely describe this procedure and we

propose to add an extra sub section (2.2.3 Spatial adjustments for the reference product), with the following text:

"The adjustment procedure to derive R_A consists of three steps: (1) mean field bias correction (one adjustment factor for the whole country which varies per hour, see Sec. 2.2.1), (2) derivation of a daily spatial adjustment factor per grid cell, and (3) spatial adjustment of the hourly or higher frequency MFB-adjusted rainfall fields (step 1) using the spatial adjustment from step 2.

A spatial adjustment factor (step 2) is derived per grid cell as follows (for a more elaborate description, see Sec. 3 in Overeem et al., 2009b):

$$F_S(i,j) = \frac{\sum_{n=1}^{N} w_n(i,j) * G(i_n j_n)}{\sum_{n=1}^{N} w_n(i,j) * R_U(i_n j_n)},$$

with N the number of radar-gauge pairs, $G(i_n j_n)$ the daily rainfall sum for manual rain gauge n at location (i_n, j_n) and $R_U(i_n j_n)$ the unadjusted daily rainfall sum for the corresponding radar grid cell. $w_n(i,j)$ is a weight for gauge n, based on the following function:

$$w_n(i,j) = e^{-\frac{d_n^2(i,j)}{\sigma^2}}.$$

Here, $d_n^2(i,j)$ is the squared distance between gauge n and the grid cell for which the factor is derived. σ determines the smoothness of the adjustment factor field. It was set to 12 km by Overeem et al. (2009b), based on the average gauge spacing in the Netherlands.

Finally, to spatially adjust the hourly MFB-adjusted rainfall fields (step 3), two more steps are followed. First, the hourly MFB-adjusted rainfall fields (see Sec. 2.2.1 for the MFB adjustment method) are accumulated to daily sums. For each grid cell, a new adjustment field is then determined:

$$F_{MFBS}(i,j) = \frac{R_S(i,j)}{R_{MFB}(i,j)},$$

with $R_S(i,j)$ the spatially-adjusted daily sum for grid cell (i,j) and $MFB(i,j)$ the MFB-adjusted daily sum for grid cell (i,j). Second, the 1-h or higher frequency (5-min in this study) MFB-adjusted rainfall fields are multiplied with adjustment factor $F_{MFBS}(i,j)$. "

Line 93 - "even": Why "even"?

Normally the unadjusted QPE (R_U) underestimates the true rainfall amount. For the first half year of 2008 this was not the case. In fact, the QPE (even) overestimates the true rainfall amount frequently.

Line 99 - "with": by

Thanks, we will change this.

Line 138 - "Delft-FEWS system": ?

Reviewer #2 also asked to give a little more information about this system. We propose to change the sentence "Most of the involved water authorities use these (lowland) rainfall-runoff models either operationally or for research purposes, often embedded in a Delft-FEWS system (Werner et

al., 2013).” into “Most of the involved water authorities use these (lowland) rainfall-runoff models either operationally or for research purposes, often embedded in a Delft-FEWS system, which is a data-integration platform, used world-wide by many hydrological forecasting agencies and water management organizations, that brings data handling and model integration together for operational forecasting (Werner et al., 2013).”

Line 139 - “For this reason, most models were already calibrated”: calibrated to what input data? Does this have an impact on the results?

This was mentioned by multiple reviewers, thanks for pointing this out. We agree that we should better clarify this procedure. Most models, except for the catchments Roggelsebeek and Dwarsdiep, were already calibrated and are part of the operational systems of the involved water authorities. Calibration took place, in most cases, with local rain gauge data for a short period of one to a couple of years. The actual calibrations of the systems generally took place not that long ago – it is different per catchment - and uses a subset of the time period used in this study (2009 – 2018). The catchments Roggelsebeek and Dwarsdiep were calibrated with the reference data (R_A) for the periods 2013 – 2014 (Roggelsebeek) and 2016 – 2017 (Dwarsdiep). The choice for these periods was based on discharge observation availability and quality.

In the validation procedure, we are using the model runs with the reference data (R_A) as ‘observation’. Hence, in any case, this validation setup will favor the model runs that are fed by QPE products that are closer to the reference rainfall product.

We propose to change “For this reason, most models were already calibrated (e.g. Brauer et al., 2014b; Sun et al., 2020).” into “For this reason, most models were already calibrated using interpolated rain gauge data or the R_A product (e.g. Brauer et al., 2014b; Sun et al., 2020). The calibration period was based on the availability and quality of discharge observations for that basin, but it was generally one to two years within the period considered in this study (2009 – 2018). The WALRUS models for catchments Roggelsebeek and Dwarsdiep were not calibrated prior to this study and were therefore calibrated with the reference data (R_A) for the periods 2013 – 2014 (Roggelsebeek) and 2016 – 2017 (Dwarsdiep). The choice for these periods was based on discharge observation availability and quality.”

Line 144 - “Kling-Gupta Efficiency (KGE)”: Please explain briefly the idea of the score. What is a good score - the higher the better?

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 Pearson correlation between observed and simulated discharge, α the flow variability error between observed and simulated discharge and β the bias factor between mean simulated (μ_s) and mean observed (μ_o) discharge. σ_s and σ_o are the standard deviation of the simulated and observed discharge. The KGE metric ranges from -∞ to 1.0, with 1.0 representing a perfect

agreement between observations and simulations. In this study, the discharge simulated with R_A as input was regarded as the observation."

Line 152 - "has one of the highest biases": In the discharge or the QPE itself?

Both actually, but the bias in the discharge simulations is a result of the bias in the QPE.

Line 139 – title: Better: Seasonal and spatial variability

Good suggestion, we will change the title to Seasonal and spatial variability.

Line 162 – "south and east": South and East (please correct all appearances in the text)

Thanks for mentioning this. We looked this up and it should only be capitalized when it is part of the name (e.g. South Africa), but for wind directions it should not be capitalized.

Line 171 – degree symbol: Replace by K (Kelvin)

We have changed it for the 5.5 K km^{-1} indications, Kelvin is indeed better.

Line 175 – 120 km: What is the range of the radars? The figure suggests 100 km?

The range is longer than the indicated 100 km. In the radar domain, a maximum range of 200 km around the radar in De Bilt was used (hence, that is what the domain is based on), so more than the entire land surface of the Netherlands is covered by the radar domain. Note that the maximum range of each radar is 320 km. However, in literature the 100 km range is often used as an indication of the maximum distance up to where the QPE is expected to be reliable. We decided to follow this 'standard' with the hope that it would not confuse the reader.

To avoid confusion for the readers, we have extended the sentence "The two grey circles indicate a range of 100 km around the radars in Den Helder (DH) and Herwijnen (H)." in the figure caption with: "The two grey circles indicate a range of 100 km around the radars in Den Helder (DH) and Herwijnen (H). Note that the used range in the composite was more than 100 km, but 100 km is often regarded as the distance up to where the radar QPE is expected to be reliable."

Line 177 – dependence: Change to 'dependency'.

We have changed it, thanks.

Line 178 – Section 3.2: The text seems to refer to mean values that are not presented in Fig. 5. Please add the mean values (including the spread) to the Figure and/or clearly indicate, what you're referring to.

Thanks for pointing this out, this was indeed missing. Also in line with the suggestions of reviewer #2, we decided to adjust the figure and add another subfigure with the annual mean absolute error between the QPE product and the reference (R_A) per catchment. This changes the figure caption as well as the text in Sec. 3.2. The proposed changes are (in addition, note that we have added more to this section after the comments of reviewer #4. We refer to our responses to this reviewer for the full adjustment and added text to this section):

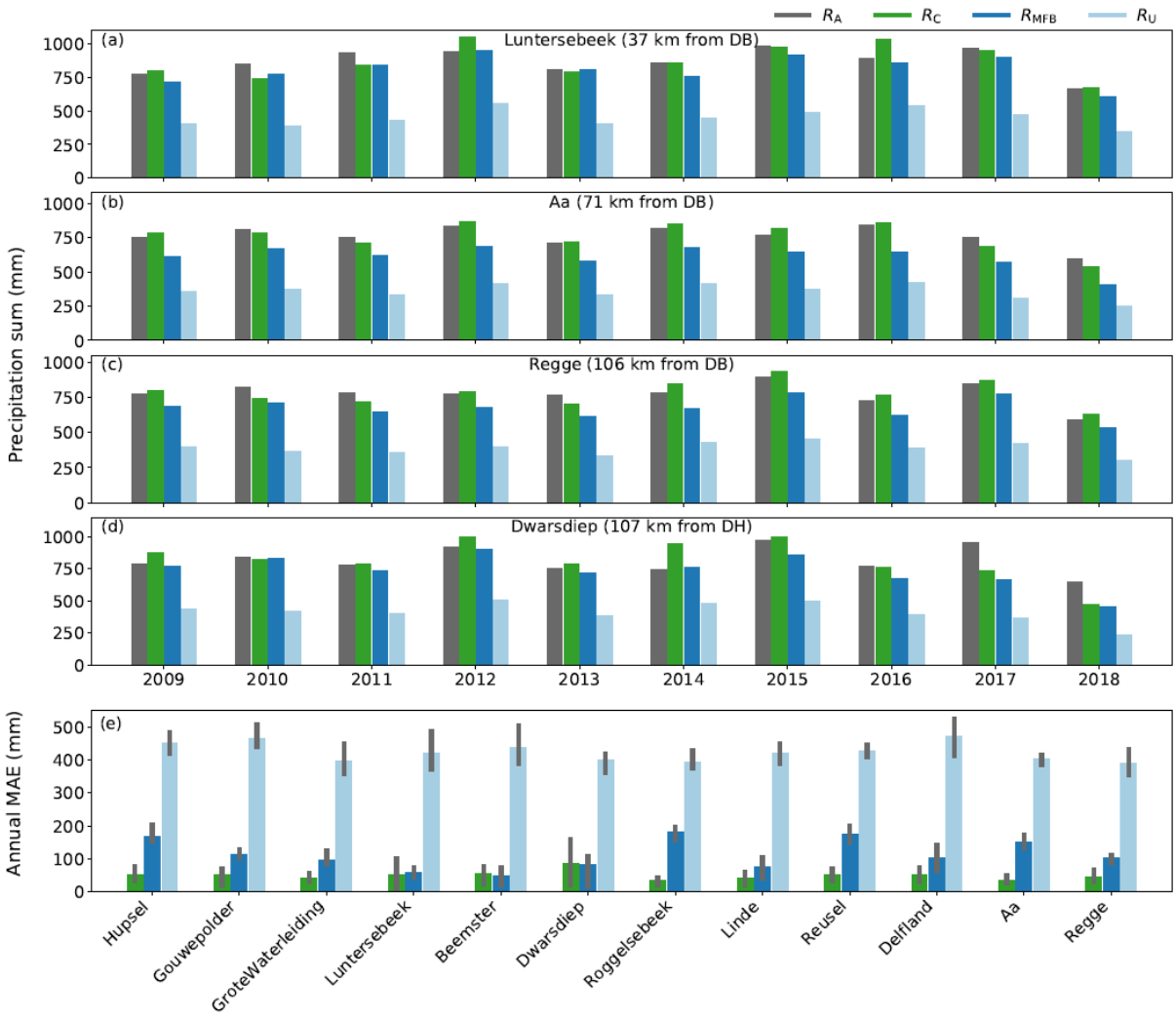


Figure 6. Effect of the adjustment factors on the catchment-averaged annual rainfall sums. (a – d) The results for a sample of four catchments that are spread over the country (and thus the radar domain): (a) Luntersebeek, (b) Aa, (c) Regge and (d) Dwarsdiep. Shown are R_A (grey), the estimated rainfall 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 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.

3.2 Annual rainfall sums

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 similar.

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 as 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. with 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 Luntersebeek catchment and Dwarsdiep polder (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 correct 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.

Line 187 – “better for the Dwarsdiep polder (10% underestimation) and Luntersebeek catchment (6% underestimation)”: This does not hold for all the years. Does it refer to the mean?

This indeed referred to the mean. After the changes to Fig. 5 (see above), the mean is shown including the spread. We propose to change this line to: “better for the Dwarsdiep polder (on average 10% underestimation) and Luntersebeek catchment (on average 6% underestimation)”.

Line 196 – “for regions close to the edges of the radar domain”: not true for Northern NL.

That is indeed not true for northern NL, we will change the sentence to: “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.”

Line 213 – “The CARROTS QPE outperforms R_{MFB}, when this product is used as input for the twelve rainfall-runoff models”: How can it be, when QPE is worse? Better in day-to-day corrections? Please give more detailed explanation on this! In addition, the sentence is not correct as stated in the next sentences that it does not hold for Beemster. Please be more precise.

With the renewed figure 5, it becomes clear that the CARROTS QPE outperforms the MFB-adjusted QPE for most catchments. We hope that this gives part of the explanation. With regard to the sentence about the Beemster, that is correct. Reviewer 2 also suggested to change this. We propose to change the sentence as follows: “The exception to this is the Beemster polder. The Beemster is mostly fed by upward seepage, leading to a predictable baseflow for all models runs. In addition, the model is located close to an automatic weather station and is located in between both operational radars, which makes the MFB adjustment more beneficial for this region. The difference in

performance between the hydrological model simulations is small, with a KGE of 0.92 (using R_C) versus 0.96 for R_{MFB}, as compared to the reference run."

Line 234 – "generally outperformed": Be more specific - this sounds too positive to me regarding the annual variability for the annual sum at two of the shown catchments.

We propose to change the sentence into: "The method and resulting QPE product outperformed the mean field bias (MFB) adjustment, that is used operationally in the Netherlands, for catchments in the east and south of the country. When the QPE products were used as input for hydrological model runs, the method outperformed the MFB adjustment method for all but one basin."

Line 236 – "The main difference with the MFB adjustments": to?

We meant the main difference with the CARROTS method. We propose to change it to: "The main difference that distinguishes the CARROTS method from the MFB adjustment is the presence of a high-density network of (manual) rain gauges in the reference dataset, a dataset that is not available in real-time."

Lines 236 – 242: To my opinion, also the timeliness of the method makes it very valuable for a first guess in real-time that should be available much earlier than products depending on gauge data.

This is a good remark and we think something that is worth mentioning at the end of this paragraph. We propose to add the following to the end of this paragraph: "An additional advantage of the method is the real-time availability of the correction factors, which is independent of the timeliness of the rain gauge data.".

Lines 243 – 247: To my knowledge, the MFB is also applied for limited areas depending on the number of gauges. Please comment on this.

That is true, see also our response to the first general comment. We are planning to add the following sentence to the end of the paragraph: "More regionalized MFB adjustments are possible, but depend on the density and availability of the automatic gauge stations.".

Lines 248 – 251: What is the experience with De Bilt?

Good point, that is something we can comment on. As stated in Sec 2.1, "Between September 2016 and January 2017, both radars were replaced by dual-polarization radars and the radar in De Bilt ('DB' in Fig. 1) was replaced by a new one in Herwijnen ('H' in Fig. 1). The radar renewals and relocation have had a limited impact on the QPE product, mainly because the operational products are not yet (fully) using the additional information from the dual-polarization (Beekhuis and Holleman, 2008; Beekhuis and Mathijssen, 2018)." Hence, a change like this has had a minor impact on the CARROTS derivation, given the historical dataset of 10 years it is based on. We do expect that when the dual-polarization potential is fully used or when an extra radar is added to the composite in e.g. the south or east (where the biases are generally highest), this no longer holds and the factors need to be recalculated using an archive based on the new situation.

We propose to add a few sentences to this paragraph in the discussion section: "However, the proposed CARROTS method has to be recalculated for every change in the radar setup, calibration, additional post-processing steps (e.g. VPR corrections, Hazenberg et al., 2013) or final composite generation algorithm. For instance, including a new radar in the composite would require a recalculation of the adjustment factors, thereby assuming the presence of an archive of the new

composite product. This could potentially limit the usefulness of the proposed method. As mentioned in Sec. 2.1, the replacement of both Dutch radars by dual-polarization radars in combination with the replacement of the radar at location De Bilt to location Herwijnen (Fig. 1) between September 2016 and January 2017 only had a limited impact on the operational products, and thereby on the CARROTS derivation. The operational products are not yet (fully) making use of the dual-polarization potential. We expect that the factors will have to be recalculated as soon as the additional information from the dual-polarization radars is used to improve the products or when e.g. the German and Belgian radars close the Dutch border are added to the composite.

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. In the case of a new radar QPE product, it is also recommended to recalculate the archive (if possible), to make sure new CARROTS factors can be derived."

Line 257 – "computationally expensive": I don't think that this is the major limitation in real-time adjustment. Computational efficient methods with spatially non-uniform factors are also common. Please change your statements from 'black-and-white' and allow for 'grey'.

We agree. Instead of this statement, we think it is worth referring back to your earlier statement about the timeliness of the CARROTS method, which makes it independent of the arrival time of the rain gauge observations in real time.

We propose to change the sentences: "A disadvantage of geostatistical and Bayesian merging methods is that they are computationally expensive and require the real-time availability of a dense network of rain gauges. Instead, we consider the proposed climatological radar rainfall adjustment method as a benchmark for the development and testing of operational radar QPE adjustment techniques." to "A possible disadvantage of these real-time methods (MFB, geostatistical and Bayesian merging) is the dependency on the timely availability of rain gauge data, which is not the case for CARROTS. Altogether, we consider the proposed climatological radar rainfall adjustment method as a benchmark for the development and testing of operational radar QPE adjustment techniques."

Line 262 – "in time": What do you mean?

This is indeed an unnecessary addition to the sentence. We have removed these words from the sentence, as the meaning of the sentence remains the same without it.

Line 295 – "factors": annual sums?

Indeed, we did mean the annual sums obtained after correcting the QPE with the CARROTS factors. Suggested change: "This bias is almost absent for the CARROTS factors" to "This bias is almost absent for the annual rainfall sums after correction with the CARROTS factors".

Line 300 – "for hydrological applications": Does this hold for extreme events? It is not stated very precisely in the next sentences. What about overforecasting in non-extreme situations?

Thanks for mentioning this, because it indeed can depend on the type of event taking place. We propose to change the focus to reconsidering the use of MFB adjustments (in this way) operationally in the Netherlands. So, the sentence would change from “For the Netherlands, these results indicate that the operationally used MFB adjustment performs worse than the proposed climatological adjustment factor for hydrological applications.” to “For hydrological applications in the Netherlands, these results indicate that the current operational use of a country-wide MFB adjustment may be reconsidered as it often performs worse than the proposed climatological adjustment factor, which can be seen as the minimum benchmark to outperform.”

Figure 5: Would be interesting to see the sum or mean over all the years (including the spread).

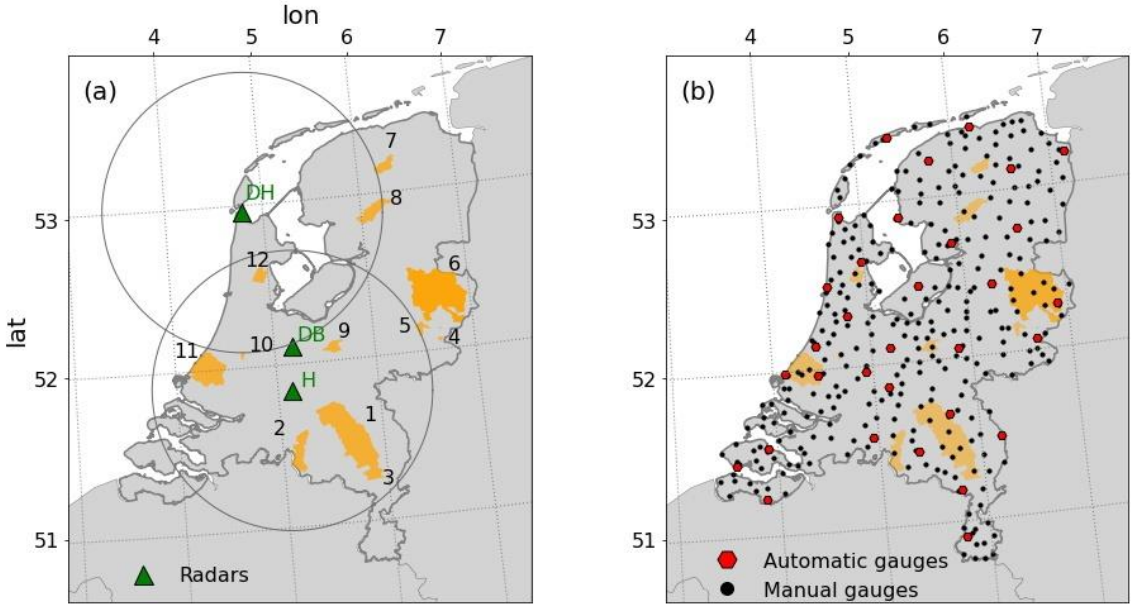
Good suggestion. See our response to Line 178 where we address this.

Author response to anonymous referee #2

General comments

1) Information about the location of the daily and hourly rain gauge should be included in Figure 1 in order to understand better the areal rainfall and discharge results. Additionally, in Table 1 rain gauges included per catchment (if applicable) can be added as an extra column. The role of the gauge density in some catchment (either hourly or daily) may explain the results of Figure 6 – for example catchment Delfand where the MFB has slightly better results.

We would like to thank the reviewer for this suggestion, it indeed directly visualizes why the MFB adjustment procedure works better for some regions than others. The new figure will look as follows (including updated caption):



(c)

Number	Name	Size (km^2)	# Automatic rain gauges	# Manual rain gauges	Used model
1	Aa	836	0	5	WALRUS
2	Reusel	176	0	1	WALRUS
3	Roggelsebeek	88	0	1	WALRUS
4	Hupsel Brook	6.5	1	1	WALRUS
5	Grote Waterleiding	40	0	0	WALRUS
6	Regge	957	1	8	WALRUS
7	Dwarsdiep	83	0	0	WALRUS
8	Linde	150	0	1	Sobek RR
9	Luntersebeek	63	0	1	WALRUS
10	Gouwepolder	10	0	1	Sobek RR
11	Delfland	379	1	4	Sobek RR
12	Beemster	71	0	1	Sobek RR-CF

Figure 1. Overview of the basins in this study: (a) study area with the location of the three radars (green triangles) operated by KNMI and the twelve basins (orange polygons). The two grey circles indicate a range of 100 km around the radars in Den Helder (DH) and Herwijnen (H). The other radar (DB) is the radar in De Bilt, which was used until January 2017 and replaced by the radar in Herwijnen; (b) locations of the 32 automatic and 319 manual rain gauges currently operated by KNMI. Note that the number of rain gauges slightly changed from 2009 until present; (c) list of the basin names, sizes, number of gauges in the basin and employed hydrological models. The numbers in the left column refer to the numbers in (a). The right column states the used model for these areas.

In addition, in the text we mention that KNMI operates 31 automatic and 325 manual rain gauges. This is more an average over the study period. For consistency with the caption of this new Fig. 1, we have replaced these numbers by 32 and 319, respectively, and we have clearly stated that this amount did change slightly over time.

2) In Figure 5 only the areal annual precipitation of 4 catchments are given and there is not enough information to understand the results of Figure 6. Instead of annual volumes for each method, another Table or Figure may be added to summarize the annual bias of Carrot, MFB and Ru for each catchment. In this case a bias equation should be given in the paper so the reader can understand the results.

Thanks for this suggestion, it makes the results clearer. Reviewer #1 suggested a similar adjustment. We decided to adjust the figure and add another subfigure with the annual mean absolute error between the QPE product and the reference (R_A) per catchment. We leave the four catchments in the figure as a highlight per year of the results and panel (e) summarizes the results for all catchments. We have decided to show the annual mean absolute error instead of a bias, because the average bias of the CARROTS QPE (over the ten years) tends to be close to 1.0 (no bias) for most catchments due to a combination of both over- and underestimations from year to year. By showing the absolute error, it is better recognizable that there are errors in the CARROTS QPE too (we think it would otherwise give a too optimistic image). The adjusted figure changes the figure caption as well as the text in Sec. 3.2. The proposed changes are (in addition, note that we have added more to this section after the comments of reviewer #4. We refer to our responses to this reviewer for the full adjustment and added text to this section):

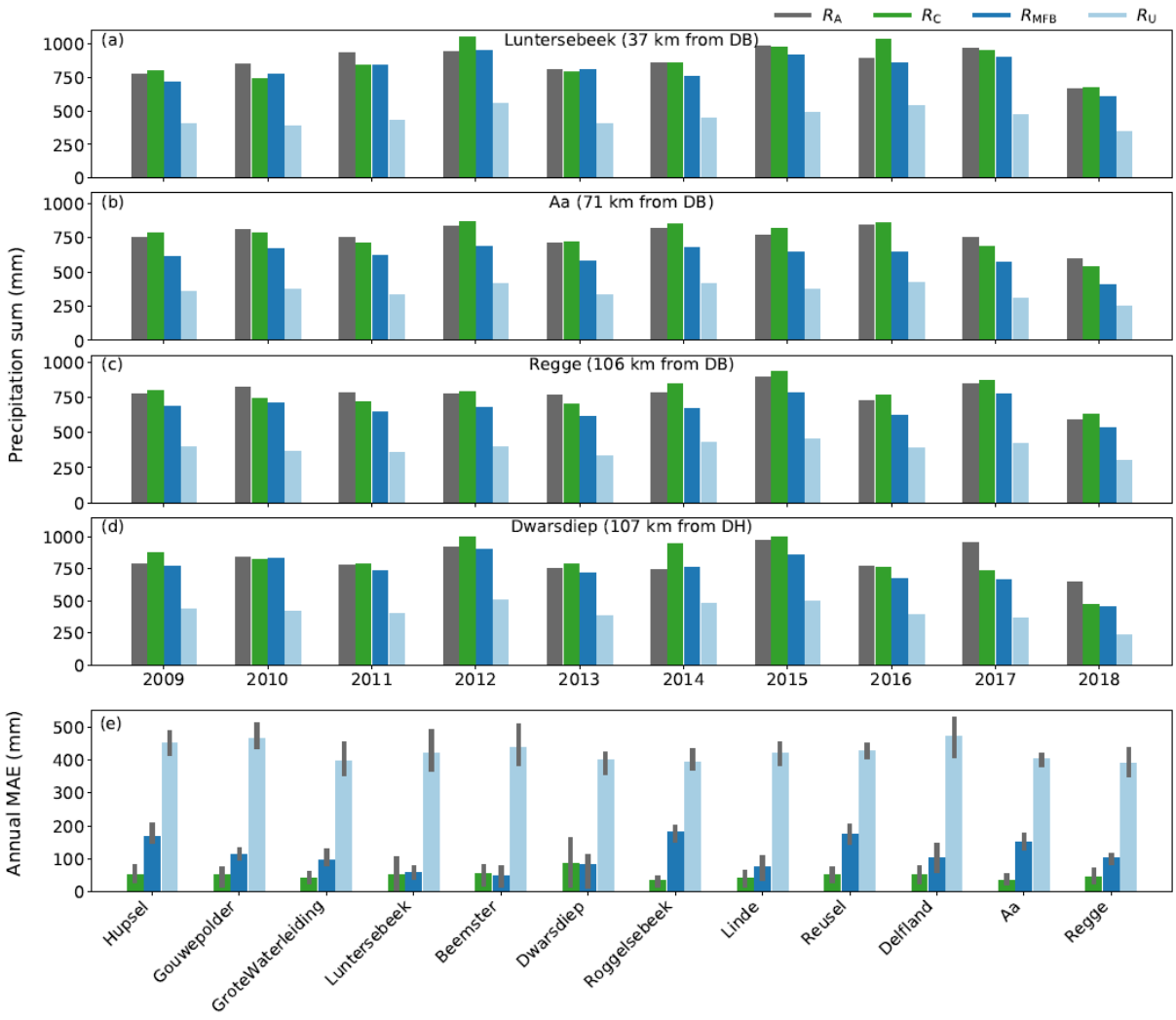


Figure 6. Effect of the adjustment factors on the catchment-averaged annual rainfall sums. (a – d) The results for a sample of four catchments that are spread over the country (and thus the radar domain): (a) Luntersebeek, (b) Aa, (c) Regge and (d) Dwarsdiep. Shown are R_A (grey), the estimated rainfall 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.

3.2 Annual rainfall sums

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 similar.

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 as 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. with 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 Luntersebeek catchment and Dwarsdiep polder (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 correct 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.

3) Another question open for discussion is the role of the catchment model type (either lumped or semi-distributed) and their calibration in the discharge errors (see comment on Figure 6). Are the semi-distributed catchments made of more than 2 sub-catchments? What data has been used for the calibration of these models?

   This was mentioned by multiple reviewers, thanks for pointing this out. We agree that we should better clarify this procedure. Most models, except for the catchments Roggelsebeek and Dwarsdiep, were already calibrated and are part of the operational systems of the involved water authorities. Calibration took place, in most cases, with local rain gauge data for a short period of one to a couple of years. The actual calibrations of the systems generally took place not that long ago – it is different per catchment - and uses a subset of the time period used in this study (2009 – 2018). The catchments Roggelsebeek and Dwarsdiep were calibrated with the reference data (RA) for the periods 2013 – 2014 (Roggelsebeek) and 2016 – 2017 (Dwarsdiep). The choice for these periods was based on discharge observation availability and quality.

   In the validation procedure, we are using the model runs with the reference data (R_A) as 'observation'. Hence, in any case, this validation setup will favor the model runs that are fed by QPE products that are closer to the reference rainfall product.

   We propose to change "For this reason, most models were already calibrated (e.g. Brauer et al., 2014b; Sun et al., 2020)." into "For this reason, most models were already calibrated using interpolated rain gauge data or the R_A product (e.g. Brauer et al., 2014b; Sun et al., 2020). The calibration period was based on the availability and quality of discharge observations for that basin, but it was generally one to two years within the period considered in this study (2009 – 2018). The WALRUS models for catchments Roggelsebeek and Dwarsdiep were not calibrated prior to this study and were therefore calibrated with the reference data (RA) for the periods 2013 – 2014 (Roggelsebeek) and 2016 – 2017

(Dwarsdiep). The choice for these periods was based on discharge observation availability and quality."

In addition, regarding the semi-distributed SOBEK RR models, these consisted of multiple sub-catchments per basin, each of which is split into paved, unpaved and greenhouse nodes (if applicable all three, but necessarily). The number of sub-catchments per basin were: 7 for Gouwepolder, 1 for Beemster, 25 for Delfland and 23 for Linde. The reference discharge was in all cases the model simulation with the reference rainfall, instead of the actual observations. In that way we have tried to be independent of any model dependencies and errors.

We propose to change lines 139 – 141 "SOBEK RR(-CF) (Stelling and Duinmeijer, 2003; Stelling and Verwey, 2006; Prinsen et al., 2010) is semi-distributed and therefore we used sub-catchment averaged rainfall sums from the gridded radar QPE." to:
"SOBEK RR(-CF) (Stelling and Duinmeijer, 2003; Stelling and Verwey, 2006; Prinsen et al., 2010) is semi-distributed and therefore we used sub-catchment averaged rainfall sums from the gridded radar QPE. The four basins that have a SOBEK model have the following number of sub-catchments: 7 for Gouwepolder, 1 for Beemster, 25 for Delfland and 23 for Linde."

4) In section 2.3 it should be stated clearly that the method is not in "leave-one-out" or split-sampling validation, the same period of data is used for the RA, for the Carrot factors, for the MFB factors and for the model calibration (or was another period used for the model calibration). Explain also shortly at this section why the Carrot factor were not used in a "leave-one-out" validation (because of low sensitivity obtained from section 3.4).

We agree that this should be explicitly mentioned. We propose to add the following paragraph to the end of Sec. 2.3:

"Note that this validation method was not a leave-one-out or split-sample validation, as the full 10-year dataset was used for R_A, the CARROTS- and MFB-adjustment derivation, and shorter periods in those 10 years were used for hydrological model calibration. However, the sensitivity of the CARROTS factor was tested by leaving individual years out of the derivation period (Sec. 2.4)."

Specific comments and technical corrections

Line 57 – "In case of a negative impact on the nowcasts, this suggests that adjustment methods should be applied to the nowcasts as a post-processing step.": What do you mean by this? That the nowcasts due to these adjustment methods are suffering from errors, and the forecaster tries to predict these errors and then adjust the nowcasts? Or by post-processing you mean adjustment after the reflectivity has been converted to rainfall rate?

We indeed mean the first, "That the nowcasts due to these adjustment methods are suffering from errors". The post-processing step would then be that the rainfall amounts are corrected after the nowcast is made, instead of prior to making the nowcast. In case a method is used that does not change the spatial structure of the rainfall fields (like MFB adjustment and CARROTS), this is not necessary. We propose to change "In case of a negative impact on the nowcasts" to "In case the nowcasts suffer from errors due to these adjustments,".

Line 84 – "both 31 automatic hourly and 325 manual daily rain gauges (Overeem et al., 2009a,b, 2011).": Spatial information about the rain gauge data would be helpful to understand the main differences between the MFB and the Carrot method. Please include in Figure 1 (or you can add

another Figure) where the hourly and the daily rain gauges are located - so that we can have a visual illustration of the station density and locations.
I think it would be nice to have two other columns here to show the number of daily and hourly rain gauges inside each catchment (if applicable) or the distance to the closest one (in case of very small catchments).

We would like to thank the reviewer for these suggestions. We have applied them. You can find our response and proposed changes under general comment #1.

Line 89 – "The R_A data is not available in real time (available with a delay of one to two months),": Could you please describe shortly how the RA is adjusted? How is the daily and hourly scaling combined together? First daily scaling and then hourly?

We agree that we can elaborate a bit more on how this method was applied. The original description can be found in Overeem et al. (2009b). We have tried to concisely describe this procedure and we propose to add an extra sub section (2.2.3 Spatial adjustments for the reference product), with the following text:

"The adjustment procedure to derive R_A consists of three steps: (1) mean field bias correction (one adjustment factor for the whole country which varies per hour, see Sec. 2.2.1), (2) derivation of a daily spatial adjustment factor per grid cell, and (3) spatial adjustment of the hourly or higher frequency MFB-adjusted rainfall fields (step 1) using the spatial adjustment from step 2.

A spatial adjustment factor (step 2) is derived per grid cell as follows (for a more elaborate description, see Sec. 3 in Overeem et al., 2009b):

$$F_S(i,j) = \frac{\sum_{n=1}^{N} w_n(i,j) * G(i_n j_n)}{\sum_{n=1}^{N} w_n(i,j) * R_U(i_n j_n)},$$

with N the number of radar-gauge pairs, $G(i_n j_n)$ the daily rainfall sum for manual rain gauge n at location (i_n, j_n) and $R_U(i_n j_n)$ the unadjusted daily rainfall sum for the corresponding radar grid cell. $w_n(i,j)$ is a weight for gauge n, based on the following function:

$$w_n(i,j) = e^{-\frac{d_n^2(i,j)}{\sigma^2}}.$$

Here, $d_n^2(i,j)$ is the squared distance between gauge n and the grid cell for which the factor is derived. σ determines the smoothness of the adjustment factor field. It was set to 12 km by Overeem et al. (2009b), based on the average gauge spacing in the Netherlands.

Finally, to spatially adjust the hourly MFB-adjusted rainfall fields (step 3), two more steps are followed. First, the hourly MFB-adjusted rainfall fields (see Sec. 2.2.1 for the MFB adjustment method) are accumulated to daily sums. For each grid cell, a new adjustment field is then determined:

$$F_{MFBS}(i,j) = \frac{R_S(i,j)}{R_{MFB}(i,j)},$$

with $R_S(i,j)$ the spatially-adjusted daily sum for grid cell (i,j) and $MFB(i,j)$ the MFB-adjusted daily sum for grid cell (i,j). Second, the 1-h or higher frequency (5-min in this study) MFB-adjusted rainfall fields are multiplied with adjustment factor $F_{MFBS}(i,j)$."

Line 136 – “validated” - Could you please explain here that no "leave-one-out" or split-sampling validation is done, because of the sensitivity analysis results.

Thanks for mentioning this. See our response to general comment #4.

Line 138 – “Delft-FEWS”: Maybe give a little bit more information about this system.

Reviewer #1 also asked to give a little more information about this system. We propose to change the sentence “Most of the involved water authorities use these (lowland) rainfall-runoff models either operationally or for research purposes, often embedded in a Delft-FEWS system (Werner et al., 2013).” into “Most of the involved water authorities use these (lowland) rainfall-runoff models either operationally or for research purposes, often embedded in a Delft-FEWS system, which is a data-integration platform, used world-wide by many hydrological forecasting agencies and water management organizations, that brings data handling and model integration together for operational forecasting (Werner et al., 2013).”

Line 139 – “calibrated”: Are the models calibrated on interpolated rain gauge data or Ra rain data? So the readers can have an idea which product the models are favouring more.

We thank the reviewer for mentioning this, as this should be better described. We have combined our answer to this point with our answer to general comment #3.

Line 144 – “Kling-Gupta Efficiency (KGE) metric”: Please explain this more and specify that the efficiency is calculated based on reference discharge (simulated from the Ra). Or do you calculate them based on real observed discharges? Also, please specify which discharge timestep do you use for the efficiency calculation.

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

Line 157: So the results of this "leave-one-year-out" validation is presented only in Figure 4-c and 7-b, right? Please mention shortly here.

That is correct, we will briefly mention the figures where the results can be found at the end of both paragraphs. In addition, we have mentioned “leave-one-year-out” in the description of the method.

Line 190 – “the MFB adjustment performs well in the north”: Is it because the gauge network is denser there?

We would like to refer to our answer to general comment 2. We have completely revised the text of that section (Sec. 3.2) and therefore this is not mentioned anymore.

Lines 215 – 216 – “The exception to this is the Beemster polder (which is mostly upward seepage driven), although the difference in performance is small, with a KGE of 0.92 (using RC) versus 0.96 for RMFB, as compared to the reference run.”: Can you please explain why do you think this catchment behaves like this? It looks quite small when compared to other catchments; is there any rain gauge that is positioned in (or very close to) the catchment? This is just an idea for discussion (also for the pros and cons of each method used).

The catchment indeed has an automatic rain gauge nearby (multiple, actually). In addition, the catchment is located in between both operational radars and therefore is located at a location in the radar domain which likely benefits most of the country-wide MFB-adjustment factor. We propose to change the sentence as follows: “The exception to this is the Beemster polder. The Beemster is mostly fed by upward seepage, leading to a predictable baseflow for all models runs. In addition, the catchment is located close to an automatic weather station and is located in between both operational radars, which makes the MFB adjustment more beneficial for this region. The difference in performance between the hydrological model simulations is small, with a KGE of 0.92 (using R_C) versus 0.96 for R_{MFB}, as compared to the reference run.”

Lines 299 – 300 – “For the Netherlands, these results indicate that the operationally used MFB adjustment performs worse than the proposed climatological adjustment factor for hydrological applications.”: Depending on how the catchments are calibrated, I would say that another advantage of the Carrot is that daily rain gauges are used in Carrot and as well in the reference product for the calibration. So they produce areal rainfall more similar to the calibration input.

We agree with the reviewer that this is an additional explanation. However, as the calibration of the catchments was different per catchment (see also our response to general comment #3), this only holds for some catchments.

Figure 5: Why are you showing here only 4 catchments? It is important to include as well the bias of the other catchments so that we can understand the behaviour of the KGE values. To simplify the results, maybe you can add another Figure or Table showing the yearly bias (for each year or average over all years) of the Carrot, MFB and Ru compared to Ra.

Thanks for this suggestion. Our explanations are given in our response to general comment 2.

Figure 6: Just a thought: it looks like for semi distributed models (j, k, l) the KGE of the Ru are better than the ones from the lumped model (except for Linde-h that is far away from the radar). This difference can be attributed also to the distance from Radar, but the KGE of other catchments in the vicinity of the radar are very low (Luntersebeek, Reusel). Do you think the choice of hydrological model may play a role here?
Another reason can be the presence of daily gauges inside catchments (which will favour more the Carrot as the Ra may be influenced more by daily gauges). Could you please discuss also the presence of daily or hourly gauges in your catchment results?

Thanks for mentioning this. The catchments that were modelled with the semi-distributed SOBEK RR model are all polders or partly polder systems. These systems are highly regulated, have high

groundwater levels and have in many cases upward seepage. Hence, the catchment behavior is somewhat more predictable for these polder systems, leading to overall higher KGE values. Note that turning on or off the pumps (e.g. in the Gouwepolder) leads to a very peaky behavior that is still challenging to simulate well. We hope that our explanations in lines 201 – 206 of the results answer this question too: "The effect is most pronounced for the freely draining catchments in the east and south of the country. These catchments are more driven by groundwater flow than the polders in the west of the country. Groundwater flow gets hardly replenished, because of similar estimated annual evapotranspiration and RU sums, resulting in too low baseflows. The polders, especially Delfland and Beemster, are an exception to this, because they are less driven by groundwater-fed baseflow and more by direct runoff from greenhouses or upward seepage flows, which makes them more responsive to individual rainfall events leading to higher KGE values (with RU as input) compared to the other basins."

Regarding mentioning the presence of daily or hourly gauges, this is a good point. We propose to add this to the end of the last paragraph in the Discussion section (instead of in the results as proposed by the reviewer):
"Finally, the CARROTS factors were derived with the reference rainfall data for the Netherlands. The same data was used as reference in this study. Although the use of the same data as training and validation set is sub-optimal, we have shown that leaving out individual years has had a limited impact on the estimated adjustment factors and the resulting QPE and discharge simulations (see also the vertical bars in Fig. 4c). Note, however, that in basins with a large number of manual rain gauges, but where automatic rain gauges are not nearby, the CARROTS results will likely be closer to the reference than the MFB-adjusted simulations. Although this is warranted for the CARROTS method, it can partly explain why the method works better for some catchments than others."

Author response to referee #3 (Søren Thorndahl)

General comments

1) In another study that the authors also refer (Schleiss 2020), we have seen significant differences between radar products based on single radars versus products based on composites of multiple radars. The latter being more reliable especially for the estimation of high-intensity rainfall. Can you maybe comment on how radar data from the two radars that you apply in the study are merged and how this relates to the larger biases as a function of distance from the radar. If there are significant range effects, I guess that this will be present both in Ra and Ru and therefore not necessarily lead to bias differences. Is there differences in the bias estimations depending on whether the location in question is covered by one or two radars.

That finding in Schleiss et al. (2020) is indeed a very interesting one. As mentioned in Sec. 2.1, the data of both radars in this study were merged using range-weighted compositing. This is further described in Overeem et al. (2009). In short, the weight a radar gets for grid cell (*i,j*) depends on the range from that radar (in km):

$$w(r) = \begin{cases} (\frac{r}{70})^2, & if\ r \leq 70 \\ 1 - \left(\frac{r-70}{200-70}\right)^2, & if\ r > 70\ and\ r < 200 \\ 0, & if\ r \geq 200 \end{cases}$$

Here, 200 km is the maximum used range of the weather radars (note that 320 km is used operationally nowadays) and 70 km is seen as the range where the weight is maximum (note that the distance between both radars is approximately 100 km, approximately 120 km between the new radars in Herwijnen and Den Helder).

The subsequent bias in this merged product is indeed influenced by the presence of multiple (two) radars in the composite. See for example also Fig. 9 in Holleman (2007), which shows that the bias is lowest in a close region around a line between both operational radars (three notes: the QPE product was slightly different then and based on a pseudo-CAPPI at 800 m instead of 1500 m now; the weighing function was slightly different in Holleman, 2007, and the color scheme for the biases is not ideal to distinguish between low ranges of biases, e.g. 0.0 to 0.6). The bias increases further away from this line, with highest errors in the East, South East and South West of the country. This corresponds also to our findings with the remaining bias in these regions after the MFB-adjustment has been applied.

R_A is also spatially corrected with the daily manual rain gauge observations, which should correct for a large part of the range effect. So, we expect that the remaining bias is low and shows up when the unadjusted radar QPE is compared to R_A.

2) The idea of a climatological bias factor works if the drop size distribution and the Z-R relationship do not change with season. It is especially important here to distinguish between convective and stratiform precipitation. In NL the winter is probably dominated by statiform precip., but summer precipitation is probably a combination of both…which would lead to more uncertainty in estimation of summer rainfall. Can you maybe provide an insight into the variability of Fclim depending on season and year. Like in figure 4 (a) where you

show MFB for 2018, but Fclim for all years with confidence bands or monthly boxplots describing the variability for each month and year.

This is a good point, the error in the summer precipitation estimation depends on the type of rainfall for a given event. CARROTS should partly compensate for differences in the Z-R relationship over the year, but for individual events this can indeed still go wrong. However, Reviewer #1 also suggested (to comment on) a dBZ-dependent correction factor. This is outside the scope of our current work, but we think it is worth looking into as a follow up. We actually performed a small analysis of the relationship between the drop size distribution and the correction factor (prior to submitting this manuscript), but this gave no clear relationship so far. However, we would like to explore this further in future work.

As F_{clim} is derived as an average over 10 years of radar data, we cannot show the annual variability of the factor. However, the sensitivity to leaving individual years out of the derivation is visualized as the vertical bars in Fig. 4c. An indication of the variability from season to season is present in Fig. 4a where a clear seasonality is present. To get an idea of the variability of the unadjusted radar QPE (R_U) quality over time, it is more insightful to plot the difference between the reference and R_U when both are accumulated over a 31-day window (the used moving window in this study).

Specific comments and technical corrections

3) In figure 7 (a) is the bias correction factor derived for all years or only 2015? Since the idea of Carrot is to use the climatological average, I would prefer to see data for the whole period 2009-2018.

   We have derived the factor for all years, just like the other results. As the factor does not change over the years (it is based on the mean over the ten years), we only show one year. In line with Fig. 6, we also only show the hydrological model simulations for the year 2015. We have chosen to show one year in order to be able to see individual discharge peaks, which would become more difficult when showing ten full years. We have stated this more explicitly in the caption to avoid any confusion: “Similar to Fig. 7, the CARROTS factors were derived for and discharge was simulated for the full period (2009--2018), but only 2015 is shown here.”

4) To be consistent please indicate on figures 3, 5 and 7 that the “bias correction factor” corresponds to Fclim.

   Thanks for suggesting this, that would make the results indeed more consistent. We have applied it to the indicated figures.

Author response to referee #4 (Marco Gabella)

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 at 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 was performed as well. The results of this analysis can be found in the appendix and the analysis was done as follows: for every rainy hour (when the sum of at least one grid cell was larger than 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_U to R_C and R_{MFB}.

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

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

"3.2 Evaluation of the rainfall sums

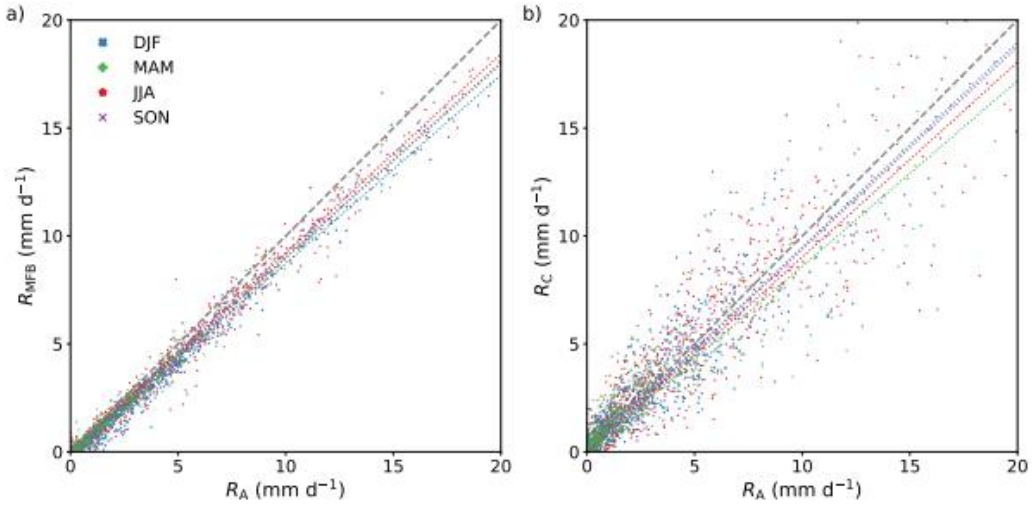


Figure 5. Comparison between the reference rainfall (R_A) and the two adjusted radar QPE products: (a) R_{MFB} and (b) R_C). Shown are the daily country-average rainfall sums based on ten years (2009–2018), classified per season. The slope, Pearson correlation and sample size per season are indicated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

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; the colored dashed lines in Fig. 5). 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). However, the remaining bias in R_{MFB} is generally caused by a systematic underestimation of the reference rainfall. The overall underestimation is less for R_C (8%, Fig. 5b), but results from estimation errors associated with either under- or overestimates of the reference rainfall. The spread in Fig. 5b is significantly wider 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 similar. 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 (mean annual R_{MFB} is 610 mm and mean annual R_A = 761 mm) and 13% for the Regge (mean annual R_{MFB} is 673 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 Luntersebeek catchment and Dwarsdiep polder (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."

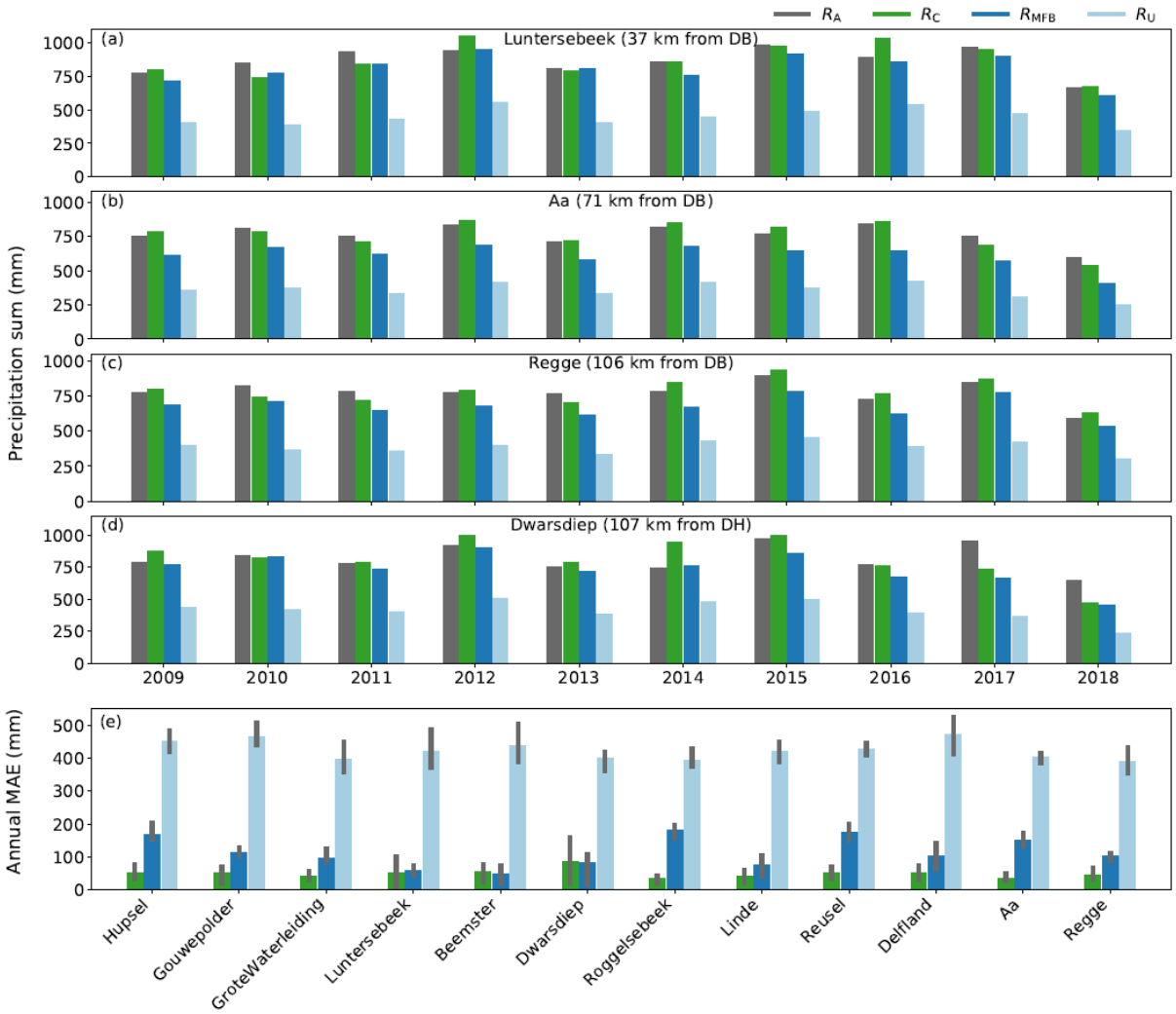


Figure 6. Effect of the adjustment factors on the catchment-averaged annual rainfall sums. (a – d) The results for a sample of four catchments that are spread over the country (and thus the radar domain): (a) Luntersebeek, (b) Aa, (c) Regge and (d) Dwarsdiep. Shown are R_A (grey), the estimated rainfall 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 have added 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 have added: “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:

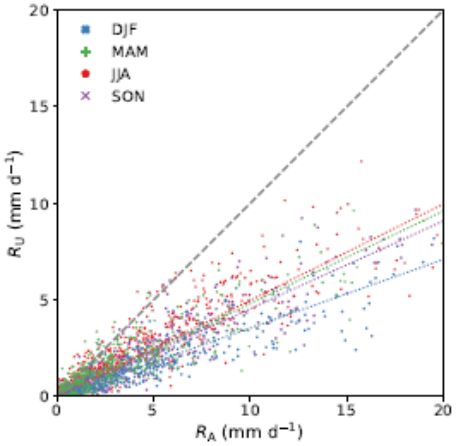


Figure 2. Scatter plot indicating the systematic discrepancy between the reference rainfall (R_A) and the unadjusted radar QPE (R_U). Shown are the daily country-average rainfall sums based on ten years (2009–2018), classified per season. The slope, Pearson correlation and sample size per season are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

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; the colored dashed lines in Fig. 2) for all observations and the Pearson correlation coefficient. This is indicated per season (DJF is winter, MAM is spring, JJA is summer and SON is autumn) 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 have added the suggested references to indicated lines.

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 now refer to Sec. 2.4 in Sec. 2.2 to indicate the presence of a sensitivity analysis: "Sections 2.4 and 3.4 describe a leave-on-year-out validation of the method and they describe the sensitivity of the method to the moving window size."

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):"

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

with

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

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

Here, r is the Pearson 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 of the simulated and observed discharge. The KGE metric ranges from -∞ to 1.0, with 1.0 representing 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 33, it would be nice to refer to Germann and Joss, 2002.

Thanks for the suggestion, we have done 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. In the case of a new radar QPE product, it is also recommended to recalculate the archive (if possible), to make sure new CARROTS factors can be derived."

Regarding elaborating on the moving window size analysis, we have extended 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. 8a). 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. 8a) lead to minimal variations in the simulated discharges for the Aa catchment, when these factors are used to adjust the input QPE (Fig. 8b). 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. 7). However, the use of a window size of 1 day or, to a lesser extent, of a week clearly leads to more fluctuations in the CARROTS factor (Fig. 8a) and can therefore influence the rainfall estimation for individual events (and the factor will also be influenced by these 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 to 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

Barnes, S. L.: A technique for maximizing details in numerical weather map analysis, Journal of Applied Meteorology, 3, 396–409, 1964.

Beekhuis, H. and Holleman, I.: From pulse to product, highlights of the digital-IF upgrade of the Dutch national radar network, in: Proceedings of the Fifth European Conference on Radar in Meteorology and Hydrology (ERAD 2008), Helsinki, Finland, https://cdn.knmi.nl/system/data_center_publications/files/000/068/061/original/erad2008drup_0120.pdf?1495621011, 2008.

Beekhuis, H. and Mathijssen, T.: From pulse to product, Highlights of the upgrade project of the Dutch national weather radar network, in: 10th European Conference on Radar in Meteorology and Hydrology (ERAD 2018) : 1-6 July 2018, Ede-Wageningen, The Netherlands, edited by de Vos, L., Leijnse, H., and Uijlenhoet, R., pp. 960–965, Wageningen University & Research, Wageningen, the Netherlands, https://doi.org/10.18174/454537, 2018.

Brauer, C. C., Torfs, P. J. J. F., Teuling, A. J., and Uijlenhoet, R.: The Wageningen Lowland Runoff Simulator (WALRUS): application to the Hupsel Brook catchment and the Cabauw polder, Hydrology and Earth System Sciences, 18, 4007–4028, https://doi.org/10.5194/hess-18-4007-2014, 2014b.

Goudenhoofdt, E. and Delobbe, L.: Generation and verification of rainfall estimates from 10-Yr volumetric weather radar measurements, Journal of Hydrometeorology, 17, 1223–1242, https://doi.org/10.1175/JHM-D-15-0166.1, 2016.

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.

Harrison, D. L., Scovell, R.W., and Kitchen, M.: High-resolution precipitation estimates for hydrological uses, Proceedings of the Institution of Civil Engineers - Water Management, 162, 125–135, https://doi.org/10.1680/wama.2009.162.2.125, 2009.

Holleman, I.: Bias adjustment and long-term verification of radar-based precipitation estimates, Meteorological Applications, 14, 195–203, https://doi.org/10.1002/met.22, 2007.

KNMI: KNMI - Jaar 2008: Twaalfde warme jaar op rij, https://www.knmi.nl/nederland-nu/klimatologie/maand-en-seizoensoverzichten/2008/jaar, 2009.

Na, W. and Yoo, C.: A bias correction method for rainfall forecasts using backward storm tracking, Water, 10, 1728, https://doi.org/10.3390/w10121728, 2018.

Ochoa-Rodriguez, S., Rico-Ramirez, M., Jewell, S. A., Schellart, A. N. A., Wang, L., Onof, C., and Maksimovic, v.: Improving rainfall nowcasting and urban runoff forecasting through dynamic radar-raingauge rainfall adjustment, in: 7th International Conference on Sewer Processes & Networks, http://spiral.imperial.ac.uk/handle/10044/1/14662, 2013.

Overeem, A., Buishand, T. A., and Holleman, I.: Extreme rainfall analysis and estimation of depth-duration-frequency curves using weather radar, Water Resources Research, 45, W10424, https://doi.org/10.1029/2009WR007869, 2009a.

Overeem, A., Holleman, I., and Buishand, A.: Derivation of a 10-year radar-based climatology of rainfall, Journal of Applied Meteorology and Climatology, 48, 1448–1463, https://doi.org/10.1175/2009JAMC1954.1, 2009b.

Prinsen, G., Hakvoort, H., and Dahm, R.: Neerslag-afvoermodellering met SOBEK-RR, Stromingen, 15, 8–24, 2010.

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.

Stelling, G. S. and Duinmeijer, S. P. A.: A staggered conservative scheme for every Froude number in rapidly varied shallow water flows, International Journal for Numerical Methods in Fluids, 43, 1329–1354, https://doi.org/10.1002/fld.537, 2003.

Stelling, G. S. and Verwey, A.: Numerical flood simulation, in: Encyclopedia of Hydrological Sciences. Part 2: Hydroinformatics, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470848944.hsa025a, 2006.

Sun, Y., Bao, W., Valk, K., Brauer, C. C., Sumihar, J., and Weerts, A. H.: Improving forecast skill of lowland hydrological models using ensemble kalman filter and unscented kalman filter, Water Resources Research, 56, e2020WR027 468, https://doi.org/10.1029/2020WR027468, 2020.

Thorndahl, S., Nielsen, J. E., and Rasmussen, M. R.: Bias adjustment and advection interpolation of long-term high resolution radar rainfall series, Journal of Hydrology, 508, 214–226, https://doi.org/10.1016/j.jhydrol.2013.10.056, 2014.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., and Heynert, K.: The Delft-FEWS flow forecasting system, Environmental Modelling & Software, 40, 65–77, https://doi.org/10.1016/j.envsoft.2012.07.010, 2013.