Referee 2

This paper compares the accuracy of weather radar rainfall using data from different countries (Denmark, Netherlands, Finland and Sweden). The study focuses on the top 50 heavy rainfall events which are more relevant for urban hydrology. The results showed that 1) radar underestimates rainfall rates; 2) radar products with higher spatial/temporal resolutions agree better with observations; 3) the combination of radar measurements from overlapping radars can improve rainfall rates. Although the results are interesting for the scientific community, there are a number of issues that the authors need to address before the paper is accepted for publication:

Major comments:

1) Rain gauge data quality. The rain gauge measurements used to validate the radar observations come from different operational agencies. It is obvious that the quality of the gauge measurements is not going to be the same among the different agencies and therefore this could impact your results. There is no discussion about this in the paper.

Response: Indeed, data quality plays a big role. To address this, more details about the type of rain gauges used in each country and their associated measurement errors/uncertainties will be added to the data section (see below). Systematic biases due to wind and calibration issues are important but unfortunately, we do not have enough information to reliably estimate them on an event-by-event basis. However, some typical values can be provided based on literature. Also, it is important to remind the reviewer that basic visual quality control has been performed on the gauge and radar data for each of the 50 events. Suspicious or obviously wrong measurements were discarded during this step. Most importantly, gauges are not considered as ground truth in this study. Rather, the goal is to describe the overall discrepancies between radar and gauge measurements, combining all sources of errors (i.e., gauges, radars, algorithms, humans) as well as differences in measurement scales.

Additional information about the rain gauges used in this study:

For Finland:
The gauges used in Finland are weight scales gauges of the type OTT Pluvio2 (https://www.ott.com). Observations are made according to World Meteorological Organization (WMO) regulations (WMO-No.8, CIMO Guide) with automatic quality control tests. Suspicious or erroneous values reported by the automatic tests are double checked manually. Measurements are based on weighing the mass of the liquid precipitation in the pluviometer, which is then converted into millimeters. A wind protector is used around the gauge. The opening of the gauge is placed at a height of 1.5 meters. For more details, see https://en.ilmatieteenlaitos.fi/weather-observations

For Denmark:
Some additional details about the Danish rain gauges can be found in Madsen et al. (2017) and the following 2 reports (in Danish):
https://www.dmi.dk/fileadmin/Rapporter/TR/tr06-15.pdf
https://www.dmi.dk/fileadmin/user_upload/Rapporter/TR/2016/DMI_Report_16_3.pdf
The full network is comprised of a combination of RIMCO tipping bucket gauges, OTT Pluvio2 weighing gauges and vibrating wire load sensors of type Geonor. But for this study, only the RIMCO tipping buckets were used.

For the Netherlands:
The automatic rain gauges in the Netherlands measure the precipitation depth using the displacement of a float in a reservoir (KNMI, 2000). The 10-min data from 2003-2017 used in this study have been validated internally by KNMI using a combination of automatic and manual quality control tests.

For Sweden:
Swedish gauges are of the type GEONOR (www.geonor.no) and consist of a bucket hanging from two chains and a metal thread that is kept in oscillation by an electromagnet. The frequency of the oscillations are transferred to a weight of the hydrometeors and is summed up. During the cold season, the bucket contains anti-freezing fluid which melts snow. The top pipe is also heated to remove larger chunks of snow and ice from clogging the opening. During the warm season an oil film is applied to keep evaporation at very low amounts. The sampling frequency is 15 min.

References:
- KNMI (2000), Handbook for the Meteorological Observation, 91–110 pp, De Bilt, Netherlands. [Available at http://projects.knmi.nl/hawa/pdf/Handbook_H01_H06.pdf]
- Madsen, H., Gregersen, I.B., Rosbjerg, D., Arnbjerg-Nielsen, K. (2017) Regional frequency analysis of short duration rainfall extremes using gridded daily rainfall data as co-variate, Water Science and Technology, 75 (8), pp. 1971-1981.

2) Rain gauge network density (Fig 1). It seems that the gauge network density is playing an important role in your results and there is little discussion on this. For Denmark the gauges are mainly clustered in a particular area (around 40-60km from radar site), for Finland the gauges are further away (beyond 50km) and cover different radars, for Sweden I can only see 4-5 gauges, whereas for the Netherlands all the gauges are more or less evenly distributed between 0-100km in range from the radar sites. This again will have important consequences in your results. For instance, VPR corrections will be important at far ranges. Attenuation due to heavy rain will also play a role. I will expect the radar rainfall error to increase with range and so the results will be better (or worse) depending on the location of the rain gauge network.

Response: Indeed, this was somewhat neglected during the analyses and discussion. To clarify this point, we will perform additional analysis and add more details about the spatial distribution of rain gauges, their distances to the radars and how much of the bias could be explained by these factors.

3) Radar data quality. Every operational agency applies different corrections to the radar data. These corrections are extremely important and can help to explain some of the results. However, there is very little detail in the paper on the actual processing steps performed by each operational agency. Some corrections are discussed, but what about corrections for attenuation, VPR, partial beam blockage, etc for some of the countries. How do you ensure that the radar data have good data quality in both rain/no rain conditions? How does the operational agencies monitor the calibration of their radars (I do not mean comparisons with rain gauge observations)? Do the bias corrections include the same (or some) of the gauges that you used for your validation? If so, what are the implications? I think this section deserves a more detailed summary.

Response: We agree this is a very important issue. But there are many factors at play and unfortunately, it is impossible to address all of them in this paper. To help with the interpretation, we will add as much information as possible about each radar product and how it was derived during revision.

For Finland:

The Finnish radar product is an experimental product from the FMI OSAPOL-project, which differs from the operational product used by the FMI mainly by making a better use of dual-polarization. The product is based on the data from years 2013-2016, during which the old single-polarization radars were being replaced by C-band dual-polarization Doppler radars. The product is therefore based on data from 4-8 dual-polarization radars depending on how many were available each year. The beam width of the radar measurements is 1 degree, the bin length is 500 m and the scanning is done in Pulse Pair Processing (PPP) mode. Doppler filtering is done first in the signal processing stage, and reflectivity measurements are calibrated based on solar signals (Holleman et al., 2010). These are followed by removal of non-meteorological targets using statistical clutter maps and fuzzy-logic-based HydroClass classification by Vaisala (Chandrasekar et al., 2013). Rainfall intensity is then estimated based on radar reflectivity (Z) and specific differential phase (Kdp) for each elevation angle. The reflectivity is attenuation-corrected (Gu et al., 2011), and the Kdp is estimated using the method described in Wang et al., 2009. For hydrometeors classified as precipitation, two alternative rain rate conversions are used. For heavy rain, i.e., Kdp>0.3 and Z>30 dBZ, the R(Kdp) relation given by $R=21Kdp^{0.72}$ (Leinonen et al., 2012) is used. For low to moderate intensities (i.e., Kdp<=0.3 or Z<=30 dBZ) and radar bins where HydroClass indicates non-liquid precipitation, a fixed Z(R) relation given by $Z = 223R^{1.53}$ (Leinonen et al., 2012) is used. A pseudo-CAPPI at 500 m height is produced from the rainfall intensity estimates using 4 lowest elevation angles and inverse distance-weighted interpolation with a Gaussian weight function. Finally, a composite VPR correction map (Koistinen et al., 2014) is applied to the resulting rainfall intensity fields that are produced at 1 km^2 and 5 min spatial and temporal resolution. The OSAPOL is the only product that is not gauge-adjusted.

For Denmark:

For the Danish C-band product, we don’t have much more information than in the Thorndahl et al. (2014) paper. But we will ask the Danish Meteorological Institute DMI for more clarifications. In addition to that, two more references to He et al. (2013) and He et al. (2018) will be provided.

For the Netherlands:

The used product is a 10-year archive of 5~min precipitation depths at 1x1 km spatial resolution based on a composite of radar reflectivities from 2 C-band radars in De Bilt and Den Helder operated by the Royal Netherlands Meteorological Institute (KNMI). Note that the Netherlands recently upgraded their radars to dual-polarization. However, the dual-polarization rainfall estimates are not fully operational yet and all rainfall values used in this study were produced with the single-polarization algorithms. Also, the radar in De Bilt stopped contributing to the composite in the course of January 2017, at which point it was replaced by a new polarimetric radar in the nearby village of Herwijnen. For a detailed description of the processing chain, the reader is referred to Overeem et al. (2009). The radars used in this study were two single-polarization Selex (Gematronik) METEOR 360 AC Pulse radars with a wavelength of 5.2 cm, peak power of 365 kW, pulse repetition frequency of 250 Hz and 3-dB beamwidth of 1 degree. The scanning strategy consists of four azimuthal scans of 360 degrees at 4 elevation angles of 0.3, 1.1, 2.0, and 3.0 degrees. The data from these scans are combined into 5-min pseudo CAPPI according to the following procedure: for distances up to 60 km from the radar, only the highest elevation angle is used to reduce the risk of ground clutter and beam blockage. For distances of 15-80 km from the radar, the pseudo CAPPI is constructed by bilinear interpolation of the reflectivity values (in dBZ) of the nearest elevations below and above the 800-m height level. For distances of 80-200 km from the radar, only the reflectivity values of the lowest elevation angle are used, whereas it should be pointed out that the 800-m level only stays within the 3-dB beamwidth of the lowest elevation up to a range of about 150 km. Values beyond 200 km from the radar are ignored. Once the pseudo CAPPI have been constructed, ground clutter and anomalous-propagation are removed using the procedure of Wessels and Beek-huis (1995), also described in Holleman and Beekhuis (2005). Spurious echoes within a radius of 15 km from the radar are mitigated based on the procedure described in Holleman (2007). A fixed Z-R relation of $Z = 200R^{1.6}$ is used to convert the reflectivities in the pseudo CAPPI to rainfall rates. During the conversion, reflectivity values are capped at 55 dBZ to suppress the influence of echoes induced by hail or strong residual clutter. Because of this, the maximum rainfall rate that can be estimated with this approach is 154 mm/h. Individual rainfall estimates from the two radars are then combined into one final composite using a weighting factor as a function of range r (km) from the radar (see Eq. 6 in Overeem et al. 2009). During the compositing, accumulations close to the radar are assigned lower weights to limit the impact of bright bands and spurious echoes. The composited rainfall rates are then adjusted for bias on an hourly basis using a network of 32 automatic rain gauges at 10 min resolution and 322 manual gauges at daily resolutions following the procedure of Holleman (2007). To improve spatial consistency, an additional bias correction at daily time scale (downscaled to hourly and 10 min scales) described in Overeem et al., (2009) is applied.

For Sweden:
The reference to Norin et al. (2015) is the only good source of information we could find. Some additional sentences will be added to the text specifying that the radars are being used for real-time operational production, and therefore prone to frequent changes and re-tuning. For example, the beam width of the radars has changed over time due to hardware upgrades and the scanning strategies and filters have been updated several times. Describing all these changes for each event and radar system quickly becomes unfeasible. Therefore, we prefer to adopt a more pragmatic approach by stating that both the technical aspects of the Swedish radar systems and their operation over the years (human and algorithm) will be assessed, with the assumption that single problematic events or radars will not have too large of an influence on the overall assessment.

References:

- Gu et al. 2011. Polarimetric Attenuation Correction in Heavy Rain at C Band, JAMC, 50, p. 39-58. https://doi.org/10.1175/2010JAMC2258.1
- He, X., Sonnenborg, T.O., Refsgaard, J.C., Vejen, F., Jensen, K.H., 2013. Evaluation of the value of radar QPE data and rain gauge data for hydrological modeling, Water Resources Research, 49 (9), pp. 5989-6005, https://doi.org/10.1002/wrcr.20471
- He, X., Koch, J., Zheng, C., Bøvith, T., Jensen, K.H, 2018. Comparison of simulated spatial patterns using rain gauge and polarimetric-radar-based precipitation data in catchment hydrological modeling, Journal of Hydrometeorology, 19 (8), pp. 1273-1288, https://doi.org/10.1175/JHM-D-17-0235.1
- Holleman, I., (2007): Bias adjustment and long-term verification of radar-based precipitation estimates. Meteor. Appl., 14, p.195–203, https://doi.org/10.1002/met.22
- Holleman, I. and Beekhuis, H. (2005): Review of the KNMI clutter removal scheme. Tech. Rep. TR-284, KNMI. [Available online at http://www.knmi.nl/publications/fulltexts/tr_clutter.pdf.]
- Holleman, I., Huuskonen, A., Kurri M. and Beekhuis, H. (2010): Operational Monitoring of Weather Radar Receiving Chain Using the Sun, Journal of Atmos. and Oceanic Tech., 27(1), p. 159-166, https://doi.org/10.1175/2009JTECHA1213.1
- Overeem, A., Holleman, I. And Buishand, A. (2009): Derivation of a 10-Year Radar-Based Climatology of Rainfall, JAMC, vol.48(), p.1448-1463, https://doi.org/10.1175/2009JAMC1954.1
- Wang. Y. and Chandrasekar, V., 2009. Algorithm for Estimation of the Specific Differential Phase, JTECH, 26(12), p. 2565-2578, https://doi.org/10.1175/2009JTECHA1358.1

- Wessels, H. R. A., and Beekhuis, J. H., 1995: Stepwise procedure for suppression of anomalous ground clutter. Proc. COST-75. Weather Radar Systems, International Seminar, Brussels, Belgium, COST, EUR 16013 EN, 270–277.

4) The radars have different spatial/temporal resolutions. This is obviously a challenge when comparing the accuracy across different operational agencies. Would not be better to accumulate to the same spatial/temporal resolution (e.g. 2x2km, 15min) in order to have a fair assessment of the results? It seems to me that the different spatial resolutions have important implications in your comparisons.

Response: Actually, because the Swedish product is at 15 min resolution and the Dutch product is at 10 min resolution, the smallest common resolution (assuming we don't want to interpolate) is 2x2 km and 30 minutes. This is rather coarse compared with the lifetime of convective cells and probably of lesser interest for most readers. We therefore think it is best to work at the highest possible resolution for the main parts of the analyses.

5) The use of ARF can help to explain the discrepancies, but I suggest to compare with the method proposed by Ciach and Krajewski (1999) which actually uses the spatial correlation of the rainfall field within the radar grid resolution to separate (or explain) the variance due to the fact that gauges represent a point whereas radar rainfall is an areal measurement from the total variance (see also Bringi et al, 2011).

Response: Thank you for the suggestion. We will compare our method to that of Ciach and Krajewski (1999) during revision to quantify how much of the bias could be explained by the spatial correlation of the rainfall fields. That being said, this is a rather delicate issue and may not necessarily result in more reliable estimates of ARFs as the latter are heavily influenced by radar data quality, the area chosen for analysis and the type of model used to represent small-scale variability. In the paper by Ciach and Krajweski (1999), an exponential function with "nugget" effect was used but other functional forms may lead to different estimates. Still, this is something that we will consider very carefully during revision.

6) Although the focus of heavy rainfall is important, what about the accuracy of radar rainfall for more conventional events (implications of the different corrections for radar errors) or in no rain conditions (e.g. implications of using robust clutter schemes, etc)? Are the results still consistent with those observed during heavy rainfall?

Response: Unfortunately, this is not feasible as only a small subset of the radar data archives has been processed so far (i.e. the top 100 events for each country, of which the 50 most intense after quality control were kept). However, it is worth pointing out that the 50 top events already contain a lot of "regular" time periods with low to moderate rainfall intensities. Therefore, a lot can be said already about the expected performance during conventional events. More information about this will be added during revision (e.g., by calculating the bias values after excluding the most intense time periods).

Other comments:

Fig 1. x/y labels? is that lat/lon?

Response: It's Lat/lon. The label will be added during revision.

Line 80. There is a reference that it is worth to look at related to the impact of spatial/temporal resolution in hydrodynamic modelling (Ochoa-Rodriguez et al, 2015).

Response: OK, thanks for the suggestion. We will add the reference.

Line 155. A lot of statements not justified: "Erroneous echoes and non-meteorological targets are removed using four different techniques. The algorithm used for correcting the vertical profile of reflectivity (VPR) is the same as in the operational product."

Response: Additional information about the radar products. See comment 3 of referee 2.

Line 160. "BRDC"?

Response: That's the name of the product. BRDC stands for BALTEX Radar Data Center, and BALTEX stands for Baltic Sea Experiment. A little note will be added in the text to explain this.

Fig 5. did you accumulate to 1h? or it is 5min,15min ...and so on?

Response: Fig 5 shows the results at 5, 10 and 15 min (as indicated)

Table 2. can you include more radar specs? e.g. beamwidth, scanning rate, radome type, pulse width, etc that can affect the measurements.

Response: Yes, some additional details will be added. See comment 3 of referee 2. However, it should be pointed out that these characteristics were not always constant over time due to hardware and software changes.

Line 420. "The total accumulated rainfall amounts per event (i.e., 10-30 mm) were lower though, suggesting that the events sampled by the X-band system were rather short and localized." For x-band radars, sometimes the radar signal might be lost due to attenuation in heavy rain and without signal there is no way to apply any correction. Is this the reason for the lower rainfall amounts? I think signal lost due to rain attenuation at X-band has to be carefully taken in to account.

Response: No, the signal was never lost during these events. We can't guarantee that the attenuation correction worked well but the relatively low bias and rmse suggests that there were no major issues. The lower rainfall amounts are simply due to the fact that there are only 2 years of data.