Reviewer 1

Review report on Evaluation of high-intensity rainfall observations from personal weather stations in the Netherlands from Rombeek et al. 2024.

The authors evaluate the robustness of rainfall estimates from personal weather stations (PWSs) by comparing them to automatic weather stations in the Netherlands over six years, identifying significant underestimation of PWS rainfall, especially for extreme events. They select rain events for different aggregations and seasons and apply part of a previously published QC method. Adjustments like bias correction improved the accuracy for moderate events, but limitations persist for high-intensity rainfall, suggesting the need for dynamic calibration and additional filtering techniques. The overall quality of the manuscript and research is good and well within the scope of HESS. To enhance its depth and utility for future readers, the authors should provide a stronger justification for their choice of QC methods, or better, show some analysis in this regarding who their choice affects the results, and propose a specific bias correction factor for hourly time scales (details provided in specific comments). Otherwise I only have minor comments and would recommend the publication of this manuscript after the above issue (selected as major, but IMO a minor major) is addressed.

We would like to thank the referee for the constructive feedback on our manuscript. We highly appreciate the time and effort to read our manuscript and provide us new insights on our research. We have taken the comments into account while preparing the revised version of our manuscript, as can be seen below.

We separated the different comments, shown in *italic,* with our replies (in regular font) below. In **bold** we provide our revised text.

Major Comments

1) Influence of QC/Bias correction on performance.

a) *You already point out the importance of QC and bias correction throughout the manuscript. Therefore, my main question is, how did the choice and the parameters of the HI and FZ filter and bias correction influence the results. I miss the reasoning for the two used filters (and not using the SO filter from de Vos et al. 2019) and not the method from Bardossy et al. (2021) or other QC methods typically used for rain gauge data. I am not asking you to compare all available methods or a detailed sensitivity analysis for each parameter, but the choice of methods and parameters will have an effect over different seasons.*
One example: the FZ filter discards a value if half of its neighbors are also zero and the HI filter relies on a maximum allowed factor that a station can deviate from the surrounding ones. Winter and summer precipitation might cause a different need of these parameters.

The goal of this study is to quantify and describe the uncertainties arising from PWS rainfall estimates, specifically focusing on the most intense events, rather than comparing QC algorithms. The key advantage of the QC algorithm from De Vos et al. (2019) over QC algorithms such as developed by Bardossy at al. (2021) and Lewis et al.

(2019) is that no auxiliary data is required. This makes it particularly suitable for regions lacking access to (real-time) reference data. For that reason, we decided to use the QC algorithm from De Vos et al. (2019). As the study from De Vos et al. (2019) also concerned rainfall in the Netherlands, we decided to use the same parameter values, rather than test how the current setting works on heavy rainfall estimates. The study from De Vos et al. (2019) used one year of data, covering all seasons and consequently different weather types, to calibrate and evaluate the filter parameters based on a reference dataset (gauge-adjusted radar data). In our study, we employed two of their filters.

We made the reasoning behind using the QC algorithm from De Vos et al. (2019) more clear in the manuscript by adding in Section 3.4, lines 196-199: **"As stated by El Hachem et al. (2024), the key advantage of the QC algorithm from De Vos et al. (2019) over QC algorithms such as developed by Bardossy at al. (2021) and Lewis et al., (2019), is that no auxiliary data is required. This makes it particularly suitable for regions lacking access to (real-time) reference data. For that reason, we decided to use the QC from De Vos et al. (2019)."**

A faulty zero (FZ) is caused when no tip occurs because the tipping bucket is obstructed completely, due to a tilted rain gauge or obstructions such as leaves or insects.
The faulty zero filter uses three parameters, which are:
- the range (*d*): stations within a given distance are selected to compute the median rainfall
- nstat: minimum number of neighbouring stations required for the filter
- nint: at least for nint time intervals the median needs to be higher than zero, while the station itself reports zero rainfall.

A high influx (HI) is caused by large recorded fluxes unrelated to rainfall itself, such as sprinklers or pouring liquids through it for cleaning.
The high influx filter uses four parameters:
- the range (*d*): stations within a given distance are selected to compute the median rainfall
- nstat: minimum number of neighbouring stations required for the filter
- ϕ_A : threshold value
- ϕ_B : threshold value

A time interval of a station is flagged as having a "high influx" if the median of the neighbouring stations does not exceed ϕ_A, while the station itself records a value above ϕ_B. When the neighbouring stations observe moderate to heavy rainfall, the station is flagged when the measurement exceed median * (ϕ_B/ϕ_A).
According to De Vos et al. (2019) most rainfall observations that should be flagged by the HI filter, were very high. They tested different subsets of parameters and found that variations in ϕ_A and ϕ_B hardly affect the results. For that reason, we decided not to change these parameters.

The calibrated value of the range parameter by De Vos et al. (2019) is 10 km for both HI and FZ. This is the average decorrelation distance of rainfall at the 5-min time interval in the Netherlands (Van Leth et al., 2021, Fig. 4a). This same work shows that this value ranges from about 10 km in summer to about 50 km in winter. In our research, we limit

the effect of spatial variability of rainfall by selecting only the five closest neighbouring stations. These PWSs have an average distance to the nearest AWS of 5.36 km, which is well within the decorrelation distance of rainfall at the 5-min time scale for any season in the Netherlands (Van Leth et al., 2021). We made this more clear in our manuscript by adding a more detailed explanation of the PWSQC algorithm and the used parameter values in Section 3.4 (Quality control algorithm), lines 202-219.

The FZ and HI filters can be applied on 5-min data, making it possible to apply them in near real-time. In contrast, the station outlier filter requires at least two weeks of data (or longer, if there is insufficient precipitation in this period), and is more computationally expensive. The effect of individual station outliers is also limited by taking the average of a cluster of stations around an automatic weather station, following the 'wisdom of the crowd'-principle.
In section 3.4 lines 224-227 we added a sentence to explain why we did not use the SO filter: "**The station outlier (SO) filter requires at least two weeks of data (or longer, if there was insufficient precipitation in this period), and is computationally expensive, which is not favourable for real-time applications. In addition, by taking the average of a cluster (minimum 5, maximum 10) of stations around an automatic weather station, the effect of individual station outliers is limited in this study. This last step is different from the method suggested by De Vos et al., (2019).**"

b) *You discuss the bias correction well and it is reasonable to use a default value from a previous study. By showing the residual bias over different aggregations e.g. in Fig 8c you already indirectly give the optimal bias correction factor. You could add this as a result to the paper extending its scope a little bit. A suggestion would be to support the bias correction factor it by giving the uncertainty through a bootstrap sampling. Checking both the filtering and bias correction would allow you to further assess the robustness of the PWS rainfall rates.*
We analysed the uncertainty of the bias as suggested by the reviewer, without applying any correction factor to the data. To estimate the uncertainty of the required bias, we applied bootstrapping (1000 iterations, with replacement). These results from Fig.1 show a differences between the seasons at a given interval and, for the same season between aggregation intervals. The bias decreases over longer aggregation intervals. This suggest that there is a need for a large correction for 1h durations. However, especially for the short durations, there are several rainfall events which have a large spatial variation in rainfall, suggesting that the PWSs are not necessarily incorrect in reporting lower rainfall.

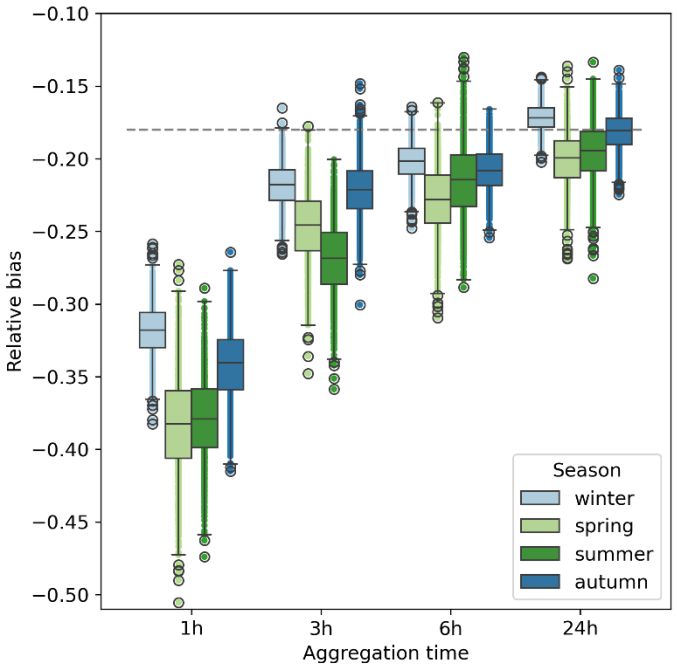

Figure 1 Estimated uncertainty in the relative bias for each season and aggregation interval based on bootstrapping (1000 iterations, with replacement). No correction factor was applied to the dataset. The lower and upper whiskers indicate the minimum and maximum relative bias and the boxes the inter-percentile range (25th -75th). The dashed horizontal grey line indicates the relative bias found by De Vos et al. (2017) based on three collocated PWSs very close to one of KNMI's AWSs.

It is important to make a distinction between these sources of bias, to avoid correcting non-instrumental related errors. The bias can be divided into two categories: 1) bias resulting from the spatial variation in rainfall extremes and 2) an instrumental bias (see Fig. 1 dashed horizontal line).

To illustrate that the bias is partially caused by category 1, we present several weather radar images to illustrate the spatial distribution of rainfall (namely, Fig. 2). Comparing rainfall accumulations at the locations of the AWS and cluster of PWSs in Fig. 2a suggests that a bias of 0.94 is present, hence a bias correction factor of 17 would have been needed to compensate for that. However, radar images show that the differences are caused by the rainfall distribution. Similarly in Fig. 2b, a bias of 0.86 is present, however, the rainfall distribution caused these differences.
The variation in relative bias in Fig. 1 are caused by the spatial rainfall variation (as shown in Fig. 2), rather than only by an instrumental bias. A higher bias correction factor would correct for the spatial variation in rainfall and not for the instrumental error.

The areal reduction factor (ARF) accounts for the spatial variability of rainfall extremes. We try to compensate for the spatial variation of the rainfall event by applying an ARF, however, such factors are based on areal rainfall climatology, representing an average behaviour and not tied to one specific event.

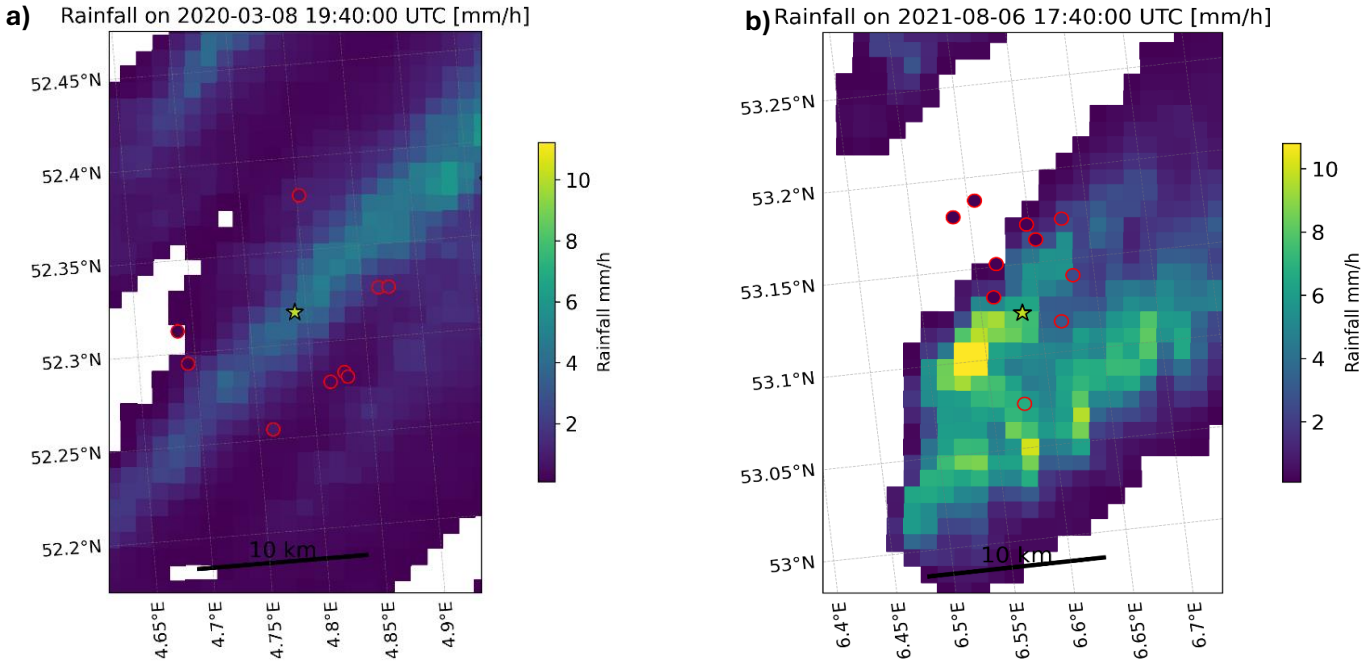

Figure 2 Rainfall distribution on a) March 8th, 2020 at 19:40 UTC and b) August 6th, 2021 at 17:40 UTC, based on 1-h accumulated rainfall from the gauge-adjusted radar product (Overeem et al., 2009). The asterisk indicates the AWS, whereas the circles with red borders represent the PWSs. The fill colour of both the asterisk and circles represents the recorded rainfall at the specific rain gauge. Comparing rainfall accumulations recorded by the AWS and PWSs indicates a relative bias of a) -0.94 and b) -0.86 present in the PWS dataset.

Based on the reviewers, we looked critically at the used default bias correction factor. In our first version, we used 1.24, based on the PWSQC from De Vos et al. (2019). This factor is a bulk correction factor tuned on gauge-adjusted radar data, and not related to individual instrumental biases, as it is not calibrated in a controlled setting. De Vos et al. (2017) used an experimental setup to investigate part of the instrument-related errors. They showed that, under ideal circumstances (i.e., compared to an electronic rain gauge installed and maintained according to World Meteorological Organization standards), three tipping-bucket rain gauges, from the Netatmo brand, recorded rainfall with high accuracy (r^2 of 0.94, based on 10-min rainfall data). However, the personal weather stations on average were found to exhibit an appreciable instrumental relative bias of -0.18 with respect to the reference gauge at the 10-min accumulation interval, suggesting the need for a default bias correction (DBC) factor of 1.22 to compensate for this systematic underestimation. We decided to use this DBC factor to adjust the instrumental bias instead of the factor of 1.24 used in our first version. We included this in Section 3.4, lines 227-230:

"As a last step of the PWSQC algorithm, a default bias correction factor (DBC) was applied to the dataset. De Vos et al. (2017) used an experimental setup and showed that under ideal circumstances there is on average an instrumental bias of 18% in the 230 Netatmo PWSs, suggesting the need for a DBC factor of 1.22 to correct these instrumental biases."

We made this distinction between the two different sources more clear in our manuscript in Section 4.3, lines 301-303:
"It is important to make a distinction in the sources of bias, to avoid correcting non-

instrumental related errors. The bias can be divided into two categories: 1) bias resulting from the spatial variation in rainfall extremes and 2) an instrumental bias."

Table 1 shows the relative bias after adjusting the values of the PWSs using the ARFs (apply the inverse of the ARFs to convert the PWS cluster to a point observation) and correcting for the instrumental bias using the factor 1.22. The remaining bias is within the expected uncertainty of rainfall observations. See Section 4.3, lines 304-315 and Table 2.

Tabel 1 Relative bias calculated after applying areal reduction factors based on Beersma et al. (2019) and correcting for the instrumental bias over the 110 (i.e.10 rainfall events x 11 AWSs) selected rainfall event per season and interval.

	Relative bias			
Interval	**DJF**	**MAM**	**JJA**	**SON**
1h	-0.02	-0.1	-0.05	-0.01
3h	0.04	-0.01	0.03	0.07
6h	0.03	-0.01	0.07	0.05
24h	0.01	-0.02	0.03	0.03

To assess the robustness of the PWS rainfall rates, we applied bootstrapping (1000 iterations, with replacement) (see Fig. 3). This figure is included in the manuscript as Fig. 9.

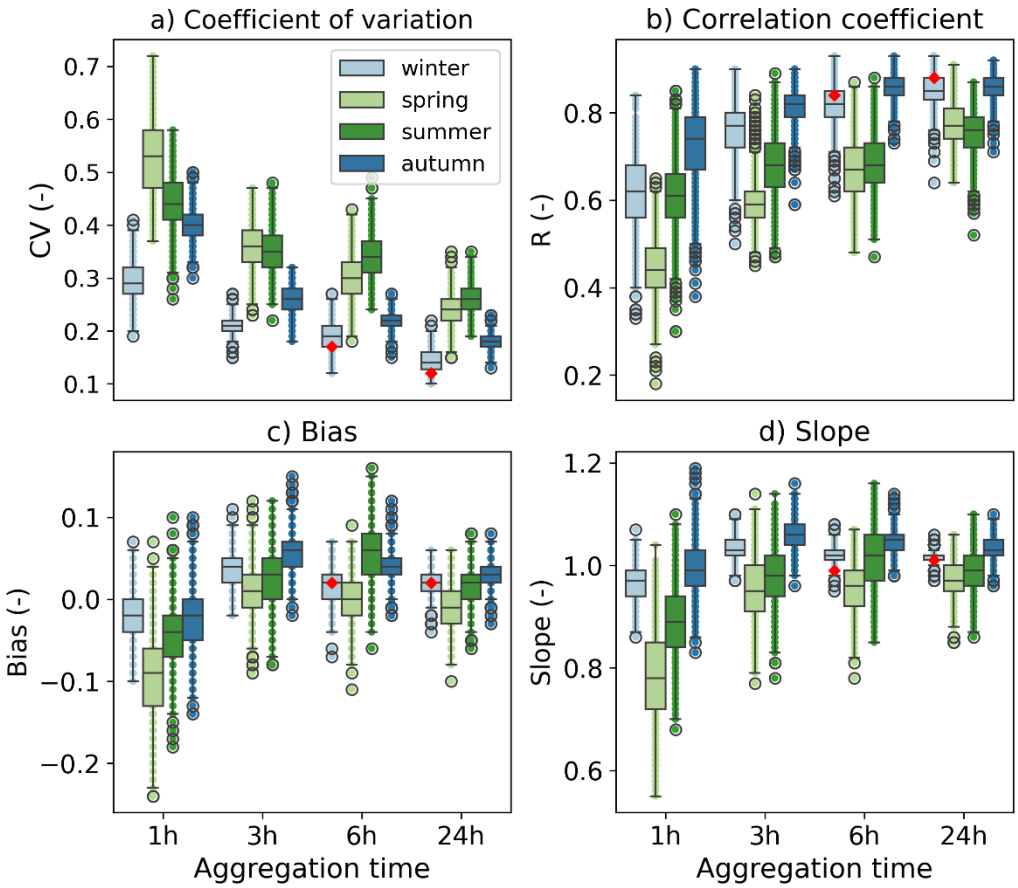

Figure 3 Coefficient of variation (a), correlation coefficient (b), bias (c) and slope (d) of filtered PWS rainfall accumulations against AWS for different seasons and accumulation intervals. The lower and upper whiskers indicate the minimum and maximum of each metrics and the boxes the inter-percentile range (25th-75th). The red diamonds indicate the values in winter after discarding events

2) Structure of motivation for this study

I find the reasoning and structure of uncertainty factors of rain gauges in general and PWS specifically given in L 85ff to be unclear and not exhaustive. Errors for rain gauges and personal weather station are for example also undercatch due to wind, solid precipitation or evaporation, which are missing already in L39. They could be a fourth group of errors in L58ff.

Also, the phrase "in addition" in L61 after a list of three points suggests that you could add a (4) item to the list?

It would be good to have a more structured and complete list of the potential errors as they motive they choice for the quality control routine. You could even link individual QC methods to errors i.e. (3) setup and maintenance à bias correction and FZ filter

This is indeed not clearly structured. We restructured this paragraph in the introduction lines 60-67, by making a distinction between PWS-related errors (setup, maintenance, rounding) and general rain gauge errors: **"Rain gauges from PWSs are prone to several sources of error. These errors can be grouped into three categories. The first category consists of PWS-related errors, such as those related to inappropriate setup and lack of maintenance of rain gauges, calibration errors, rounding due to data processing, as well as connectivity issues during data transfer (De Vos et al., 2017, 2019). The second**

category includes general rain gauge-related errors, such as undercatch due to wind, solid precipitation or evaporation, and the intrinsic tipping bucket error, resulting from the given volume of water that needs to be collected before the bucket tips (Habib et al., 2001). A third category of errors arise due to spatial sampling uncertainties resulting from gauges that are not co-located and thus differences between point rainfall estimates or estimating areal rainfall using point measurements (Villarini et al., 2008)."

3) Minor comments

a) *Abstract: You could sharpen the scope of the paper by including the content from L79 (the gap you aim to close) in it.*
We included this in the abstract, lines 6-8: **"A systematic long-term analysis involving PWS rainfall observations across different seasons, accumulation intervals and rainfall intensity classes is missing so far."**

b) *Introduction:*
You may include following literature if you find them fitting

- *https://doi.org/10.1016/j.atmosres.2024.107228 give additional motivation to evaluate the performance of PWS as they are already used in applications like radar adjustment.*
  We included this in the introduction where we mention that PWS are used to adjust radar, lines 89-90: "**Note that in previous research from Overeem et al. (2024a) and Nielsen et al. (2024) rainfall estimates from PWSs were actually used to correct a rainfall radar product."**
- *https://www.jstor.org/stable/26496995 also investigate PWS in the Netherlands, focusing e.g. on the difference between urbn and rural stations*
  This is an interesting paper, however it has a different scope so we decided not to include it in our manuscript.
- *https://doi.org/10.1145/3276774.3276792 have a bit a different perspective on QC while also relying an station clustering*
  We included this in the introduction where we mention different QC methods, lines 80-81: "**Chen et al. (2018) assigned trust scores based on spatial consistency between stations."**
- *https://doi.org/10.1016/j.ejrh.2021.100883 compared (interpolated) PWS data for different time scales*
  We included this in the introduction, lines 82-85: **"While previous work has shown that implementing these QC algorithms yields an overall improvement in the quality of the PWS data (De Vos et al., 2019; Bárdossy et al., 2021; Graf et al., 2021; Overeem et al., 2024a; Nielsen et al., 2024; El Hachem et al., 2024), a systematic long-term analysis of the QC algorithm of De Vos et al. (2019) for different seasons, accumulation intervals and intensity classes is missing so far."**

  We thank the reviewer for the suggested literature, and included those fitting the manuscript .

c) *L29 560 ha seems very specific, could you better give a range of what is considered a small, fast reacting catchment*

We made this more general in lines 29-31: "**Especially small, fast-responding catchments require accurate rainfall observations with high spatial and temporal resolution for reliable predictions, such as in the order of kilometres and minutes for catchment areas of a few square kilometres.**"

d) *L52 For Netatmo I think users can decide whether data is uploaded or not? For other platforms, that is certainly the case.*

You are correct, this is not clear from the sentence. We changed it in the introduction in lines 53-56 to: "**Once the PWS is connected to an online platform such as the Weather Observations Website (WOW; https://wow.metoffice.gov.uk/), the Weather Underground website (https://www.wunderground.com/wundermap) or Netatmo (https://weathermap.netatmo.com/), observations are automatically uploaded to the respective platform.**"

e) *L77 gives the impression that you also use the QC from Bardossy et al (2021) – which would be interesting, but might be too much here*

We changed this in the introduction, lines 82-85: "**While previous work has shown that implementing these QC algorithms yields an overall improvement in the quality of the PWS data (De Vos et al., 2019; Bárdossy et al., 2021; Graf et al., 2021; Overeem et al., 2024a; Nielsen et al., 2024; El Hachem et al., 2024), a systematic long-term analysis of the QC algorithm of De Vos et al. (2019) for different seasons, accumulation intervals and intensity classes is missing so far**"

f) *L93 data availability was too low before, right? You could state this more specifically here*

Indeed, the data availability was too low before. We changed this into (see Section 2, lines 99-100: **"This period was chosen due to the PWS data availability, which was too low (less than 5 PWSs within 10 km distance of the AWS) before 2018 and increased over the years."**

g) *L102 You could add a statement about spatial correlation from de Beek (2012) already here to describe the area further*

We added this sentence in Section 2, lines 109-111: "**These different rainfall characteristics lead to a distinct seasonal cycle in spatial rainfall correlation in the Netherlands (Van de Beek et al., 2012, Fig. 4b), with longer correlation distances for winter than summer.**"

h) *Fig 2. You could add IQR or min/max to the monthly barplots to give a feeling for variability*

We included the IQR to the barplot, see Fig. 4 below:

i) *L146 Did you use a fixed window for resampling i.e. always to the full hour like XX:00 to XX:55? This could be important for the selection of events.*

For selecting the events based on the 10-min data, we used a moving window. This 10-min data is not quality-controlled by KNMI. For that reason, we did a plausibility check with the hourly data (clock hour). As this is not clear from the paragraph, we changed this in Section 3.2, lines 161-162: "**Rainfall observations from the 10-min dataset of the AWSs was employed to make a selection of events between 2018 and 2023 using a moving window approach. Only the 10 largest rainfall accumulations were selected, as these are the most important ones for pluvial flood forecasting.**

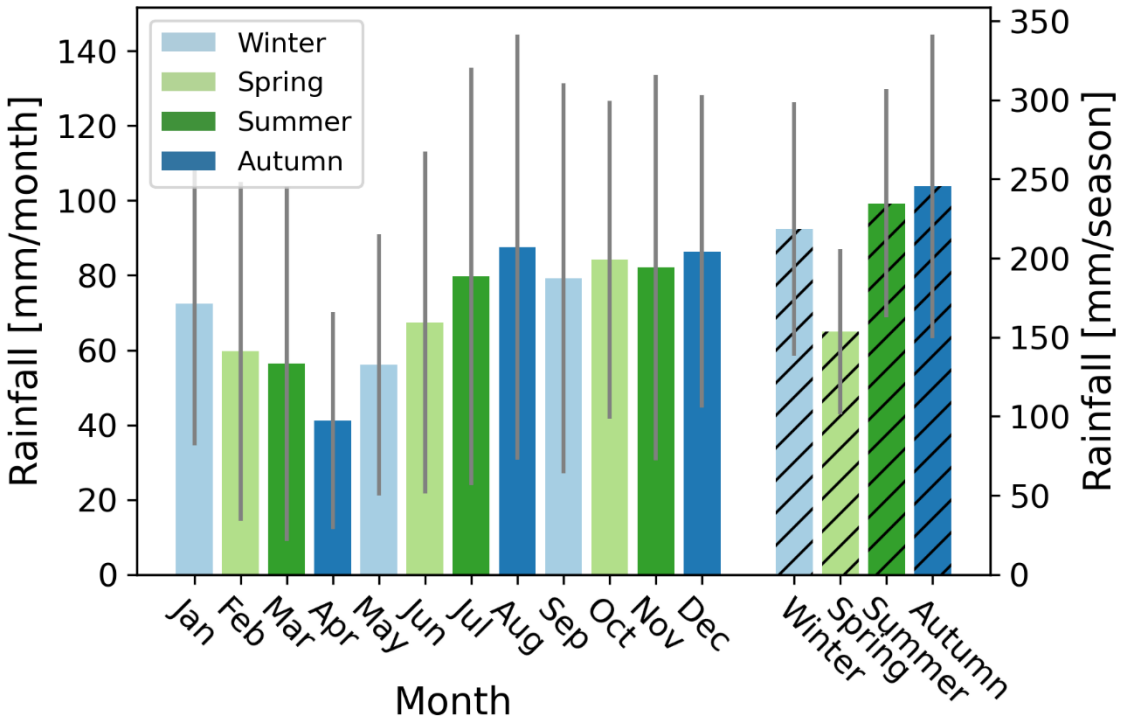

Figure 4 Average rainfall per month and season in the Netherlands over the period 1991-2020, based on data from 13 automatic weather stations spread over the country, obtained from KNMI (2024). Coloured bars indicate the average rainfall per month (mm month^{-1}), coloured hatched bars indicate the average rainfall for each season. Vertical grey lines indicate the inter quartile range.

Analysis shows that, on average, the 10 highest 1-h rainfall accumulations for the selected AWSs account for 12.5% of the annual rainfall. The hourly dataset (clock-hour) was employed to perform a consistency check on this selection."

j) *L189 You could add the best/worst/range for all validation metrics*

We included this for all metrics used, see Section 3.6.

k) *L195 You mix bias and relative bias here and, in the text, please clarify*

We changed it to relative bias (see Section 3.6)

l) *L 220 Will the*

We are not sure what is meant by this comment.

m) *L276 with an average PWS intensity*

This is indeed unclear. We changed: "The highest rainfall accumulations were observed during summer and spring, with an average intensity of more than 35 mm/h." into (see Section 4.5, lines 361-362**: "During summer and spring the highest rainfall accumulations were observed by the AWSs, with intensities exceeding 35 mm h^{-1}."**

n) *Figure 8: Any idea why JJA and MAM seem to be very similar and DJF and SON?*

As seen from Fig. 5 in the manuscript, the highest rainfall intensities occur in JJA and MAM. These highest intensities also stand out clearly in Figs. 8, A1and E1 in the manuscript, skewing the result.

o) *L345 and L353 are a bit counter-intuitive. You want to give insight in he uncertainty, but at the same try to reduce uncertainty due to aggregation. Please clarify.*

In our research the effect of individual station outliers is limited by taking the average of a cluster of stations around an AWS, following the 'wisdom of the crowd'-principle. This gives insight in systematic errors across the PWS network. We changed the text in the

conclusion, 441-442: **"This study provides insight into the systematic errors across the PWS network during high-intensity rainfall events by performing a comprehensive analysis over six years."**

We also specified that metrics are calculated over a cluster (lines 447-448): **"To reduce uncertainty from single stations, metrics were calculated over a cluster of PWSs, rather than individual stations."**

p) *L364 to 366 Do you refer to the two highest hourly events in JJA? Maybe you could check the radar images for those two events and check the spatial distribution of rainfall during these events? Similar, looking at the 5/10 minute time series from AWS and surrounding PWS could give some insight on those two events.*

This is indeed about the highest hourly events in JJA and MAM. We made this more explicit in the revised manuscript (see conclusion, lines 458-462):
"In addition, PWSs did not observe the most intense rainfall events, with high intensities over a relative short amount of time (e.g. > 75 mm h^{-1} within 10 min). These highest intensities occurred during summer and spring, with events that typically occur once in 10 years or even longer return periods. The spatial footprint of these high-intensity rainfall events is often small, influencing errors related to spatial sampling due to the average distance of 4.4 km towards the nearest AWS."

We considered the radar images of these events (see Fig. 5). The asterisk indicates the location of the AWS, which measured 67 mm in one hour, whereas circles indicate the locations of the PWSs. The differences are indeed mainly caused by the spatial

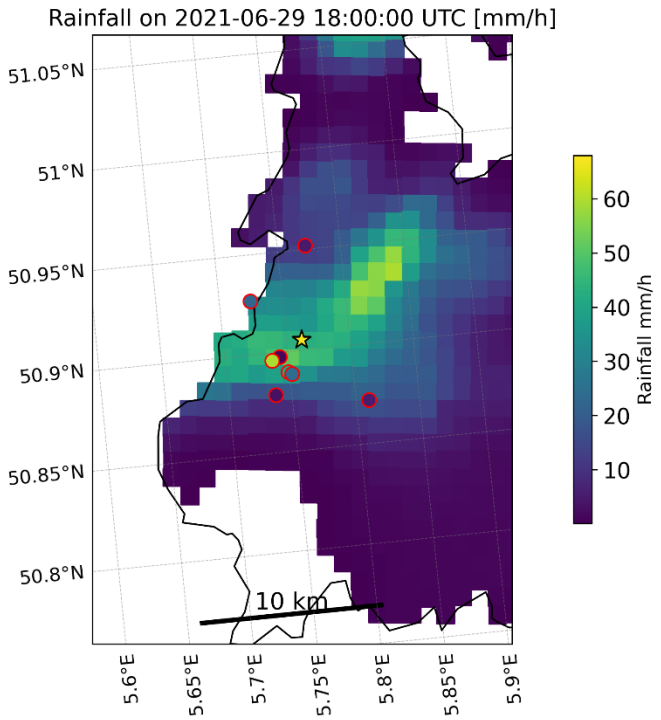

Figure 5 Rainfall distribution on June 29th, 2021 at 18:00 UTC, based on 1-h accumulated rainfall from the gauge-adjusted radar product (Overeem et al., 2009). The asterisk indicates the location of the AWS which measured 67 mm in one hour, whereas circles with red borders represent the locations of the PWSs. The fill colour of both the asterisk and circles represents the recorded rainfall at the specific rain gauge.

distribution of the rainfall relative to the locations of the PWSs. Similarly, this is the case for the other outliers in spring and summer (e.g. Fig. 6 below).

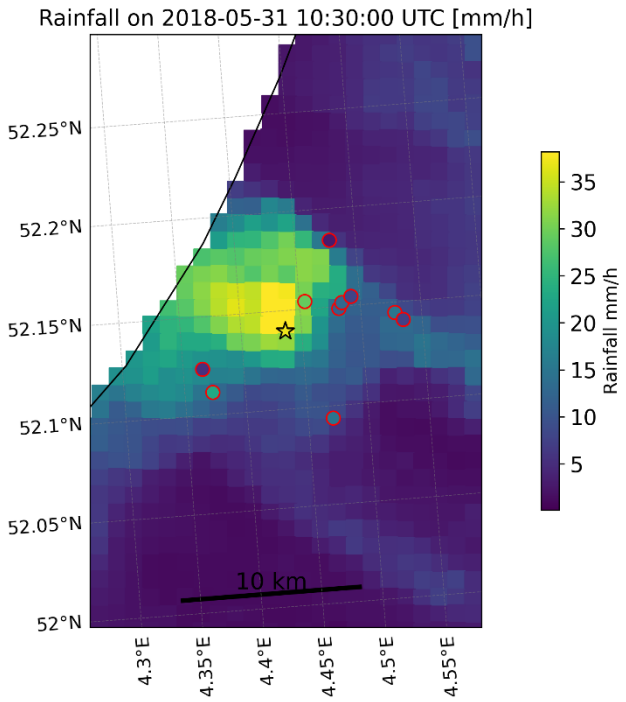

Figure 6 Rainfall distribution on May 31st, 2018 at 10:30 UTC, based on 1-h accumulated rainfall from the gauge-adjusted radar product (Overeem et al., 2009). The asterisk indicates the location of the AWS which measured 39 mm in one hour, whereas circles with red borders represent the locations of the PWSs. The fill colour of both the asterisk and circles represents the recorded rainfall at the specific rain gauge.

4) Technical issues

- *L15 duplicated "with"*
  We removed one "**with**"
- *L52 "are" instead of "is"*
  We changed it to "**are**"

Reviewer 2

In this manuscript, the authors present a systematic study of uncertainties from personal weather stations for different seasons and rainfall durations which were evaluated against a reference dataset from the Royal Netherlands Meteorological Institute. The paper is well written and the topic fits into the scope of HESS. However, the are some aspects in the research design which need to be addressed before publication.

We would like to thank the referee for the constructive feedback on our manuscript. We highly appreciate the time and effort to read our manuscript and provide us new insights on our research. We have taken the comments into account while preparing the revised version of our manuscript, as can be seen below.

We separated the different comments, shown in *italic,* with our replies (in regular font) below. In **bold** we provide our revised text.

1) *One major point of criticism from my perspective is the approach to bias correction used in this study. If the results show that the bias varies with different durations, then why is a single static value used? When you choose a MFBC of 1.24, the bias is reduced for longer durations, but this choice appears somewhat arbitrary. For instance, a higher MFBC value could potentially reduce the bias for shorter durations but might lead to an overestimation of longer durations. I miss the reasoning behind the choice of this value. Rather than applying a single MFBC value, the study could focus on providing insights into the biases of PWS with respect to duration an intensity, as already indicated in Table 2 and Figure 8 for example. Additionally, more sophisticated methods, such as quantile mapping, might be more appropriate for bias correction, as biases in PWS data can be either positive or negative. It may also be worth considering applying bias correction individually for each PWS.*

   The goal of this study is to quantify and describe the uncertainties arising from PWS rainfall estimates, specifically for the most intense events, rather than improving QC algorithms as such. We used the quality control algorithm from De Vos et al. (2019). As the study from De Vos et al. (2019) also concerned rainfall in the Netherlands, we decided to use the same parameter values, rather than test how the current setting works on heavy rainfall estimates. The last step of this PWSQC algorithm is applying a default bias correction (DBC) factor. That study reported a default bias correction factor of 1.24, calibrated using gauge-adjusted weather radar data. However, this is indeed a bulk correction factor (median of all PWS relative bias values of 5-min observations during May 2016 in the urban area of Amsterdam) and not related to individual instrumental biases, as it is based on weather radars and not in a controlled setting. Based on the reviewer, we decided to look critically at this DBC factor.

   De Vos et al. (2017) used an experimental setup to investigate part of the instrument-related errors. They showed that, under ideal circumstances (i.e., compared to an electronic rain gauge installed and maintained according to World Meteorological Organization standards), three tipping-bucket rain gauges, from the Netatmo brand, recorded rainfall with high accuracy (r^2 of 0.94, based on 10-min rainfall data). However,

the personal weather stations on average were found to exhibit an appreciable instrumental relative bias of -0.18 with respect to the reference gauge at the 10-min accumulation interval, suggesting the need for a default bias correction (DBC) factor of 1.22 to compensate for this systematic underestimation. We decided to use this DBC factor to adjust the instrumental bias instead of the factor of 1.24 used in our first version. We included this in Section 3.4, lines 227-230:

"As a last step of the PWSQC algorithm, a default bias correction factor (DBC) was applied to the dataset. De Vos et al. (2017) used an experimental setup and showed that under ideal circumstances there is on average an instrumental bias of 18% in the 230 Netatmo PWSs, suggesting the need for a DBC factor of 1.22 to correct these instrumental biases."

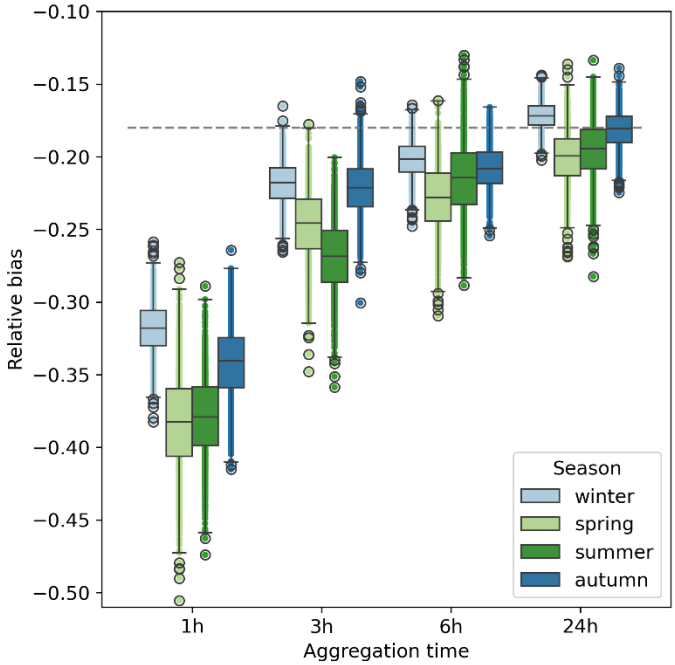

Figure 2 Estimated uncertainty in the relative bias for each season and aggregation interval based on bootstrapping (1000 iterations, with replacement). No correction was applied to the dataset. The lower and upper whiskers indicate the minimum and maximum relative bias and the boxes the inter-percentile range (25th -75th). The dashed horizontal grey line indicates the relative bias found by De Vos et al., (2017) based on three collocated PWSs very close to one of KNMI's AWSs.

We analysed the uncertainty of the bias as suggested by the reviewer, without applying any correction factor to the data. To estimate the uncertainty of the required bias, we applied bootstrapping (1000 iterations, with replacement). These results from Fig.1 show a differences between the seasons at a given interval and, for the same season between aggregation intervals. The bias decreases over longer aggregation intervals. This suggest that there is a need for a large correction for 1h durations. However, especially for the short durations, there are several rainfall events which have a large spatial variation in rainfall, suggesting that the PWSs are not necessarily incorrect in reporting lower rainfall.

It is important to make a distinction between these sources of bias, to avoid correcting non-instrumental related errors. The bias can be divided into two categories: 1) bias resulting from the spatial variation in rainfall extremes and 2) an instrumental bias (see

Fig. 1 dashed horizontal line).

To illustrate that the bias is partially caused by category 1, we present several weather radar images to illustrate the spatial distribution of rainfall (namely, Fig. 2). Comparing rainfall accumulations at the locations of the AWS and cluster of PWSs in Fig. 2a suggests that a bias of 0.94 is present, hence a bias correction factor of 17 would have been needed to compensate for that. However, radar images show that the differences are caused by the rainfall distribution. Similarly in Fig. 2b, a bias of 0.86 is present, however, the rainfall distribution caused these differences.
The variation in relative bias in Fig. 1 are caused by the spatial rainfall variation (as shown in Fig. 2), rather than only by an instrumental bias. A higher bias correction factor would correct for the spatial variation in rainfall and not for the instrumental error.

The areal reduction factor (ARF) accounts for the spatial variability of rainfall extremes. We try to compensate for the spatial variation of the rainfall event by applying an ARF, however, such factors are based on areal rainfall climatology, representing an average behaviour and not tied to one specific event.

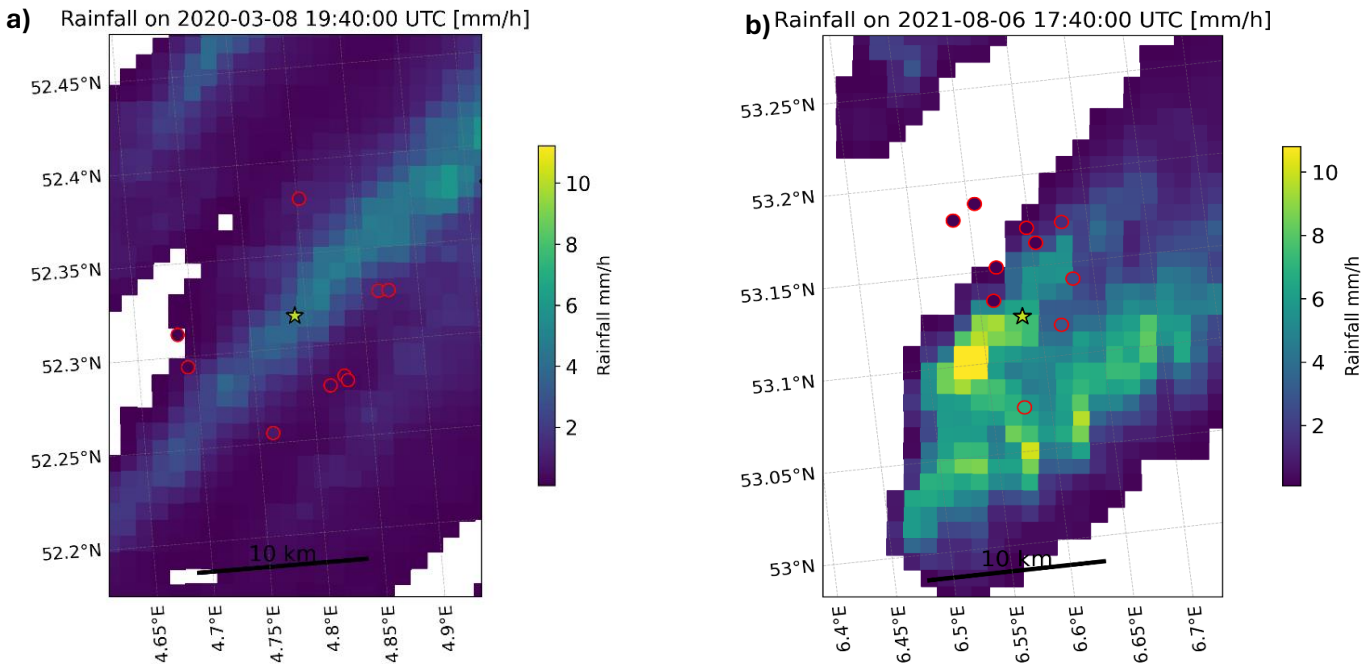

Figure 2 Rainfall distribution on a) March 8th, 2020 at 19:40 UTC and b) August 6th, 2021 at 17:40 UTC, based on 1-h accumulated rainfall from the gauge-adjusted radar product (Overeem et al., 2009). The asterisk indicates the AWS, whereas the circles with red borders represent the PWSs. The fill color of both the asterisk and circles represents the recorded rainfall at the specific rain gauge. Comparing rainfall accumulations recorded by the AWS and PWSs indicates a relative bias of a) 0.94 and b) 0.86 present in the PWS dataset.

We made this distinction between the two different sources more clear in our manuscript in section 4.3, lines 301-303:
"It is important to make a distinction in the sources of bias, to avoid correcting non-instrumental related errors. The bias can be divided into two categories: 1) bias resulting from the spatial variation in rainfall extremes and 2) an instrumental bias."

Table 1 shows the relative bias after adjusting the values of the PWSs using the ARFs (apply the inverse of the ARFs to convert the PWS cluster to a point observation) and correcting for the instrumental bias using the factor 1.22. The remaining bias is within the expected uncertainty of rainfall observations. See Section 4.3, lines 304-315 and Table 2.

Tabel 2 Relative bias calculated after applying areal reduction factors based on Beersma et al. (2019) and correcting for the instrumental bias over the 110 (i.e.10 rainfall events x 11 AWSs) selected rainfall event per season and interval.

	Relative bias			
Interval	**DJF**	**MAM**	**JJA**	**SON**
1h	-0.02	-0.1	-0.05	-0.01
3h	0.04	-0.01	0.03	0.07
6h	0.03	-0.01	0.07	0.05
24h	0.01	-0.02	0.03	0.03

We agree that more sophisticated methods might be necessary to correct for intensity-related biases and also recommended (e.g. lines 521-523) that it would be very valuable to further investigate the dynamic response of these tipping-bucket stations at different intensities to enable dynamic calibration and consequently minimize non-linear errors as a function of rain rate. However, this needs to be done separately in a controlled setting, as it is not possible to derive such a calibration from the data employed in this study.

2) *Another aspect you should consider are outliers in the winter: Why didn't you apply the temperature filter directly and exclude obvious snowfall events from the analysis? Data from unheated rain gauges, like those from Netatmo, can be inaccurate in winter, potentially leading to outliers and other issues. From my perspective, demonstrating the process of detecting these outliers and showing how results improve after their removal doesn't offer any new insights. Removing events with snowfall a priori would prevent the evaluation of obviously erroneous data, as shown in figures 7b) and 8a) and b).*

It is indeed no new insight that PWSs are unheated, which could result in inaccurate observations in winter. However, we cannot determine unambiguously whether we are dealing with solid precipitation solely by considering the temperature. By using a temperature-based flag, end users can have an indication that observations might be uncertain. We added in section 4.5, lines 354-357:
"It is not possible to unambiguously determine whether precipitation is solid based solely on temperature. However, a temperature-based flag can provide end users with an indication that the rainfall observations may be subject to uncertainty. Flags were assigned to timeseries where the corresponding AWS recorded temperatures below freezing.

3) *Last but not least, it isn't entirely clear to me what the motivation behind the ARFs is. They are only briefly addressed in section 3.6 and 4.5. Are you trying to compare ARFs from Radar with PWS? Or are using the ARFs to adjust the values of the AWS to the*

cluster size? This topic should either be motivated and discussed in more detail in the manuscript or omitted entirely.

The differences in the bias between the seasons and aggregation intervals can be partly explained by the areal reduction factor (ARF). Our approach makes use of the 'wisdom of the crowd'-principle, which refers to the collective input of individuals leading to more accurate outcomes, rather than using a single source. Using a cluster of PWSs limits the effect of outliers or errors in a single station's measurements. By averaging the rainfall observed by this cluster, however, we are effectively representing the rainfall over a broader area, while we compare it with a point measurement (the reference observation). Especially for rainfall extremes, accumulation for a given return period at a point is larger than over an area. The ARF is used to account for the fact that extreme mean rainfall over a larger area will be smaller than extreme rainfall at a single point, especially for events with a longer return period and shorter duration. The ARF partially explains the larger underestimation observed for short intervals. For that reason and based on the reviewers suggestion, we adjusted the values of the PWSs using the ARFs (apply the inverse of the ARFs to convert the PWS cluster to a point observation) before we evaluate the performance of the PWSs.

4) Minor comments

a) *Regarding the selection of the events, why didn't you select events such that the PWS data for the events have no missing values, i.e. are complete? This would also avoid the problem of missing data and underestimation as discussed in l. 240ff. If only one PWS in the cluster has missing data around the AWS infilling could also be an option.*

Approximately every 5 min data is transferred from the outdoor module (rain gauge) to the indoor module. The Netatmo processing software (which is not openly accessible) resamples this data to regular 5 min time intervals. However, careful investigation reveals that when within a 5 min interval no data is transferred this time interval is not included by Netatmo. To show this, we downloaded a sample of raw data and the aggregated data by Netatmo. "Raw" is likely the timestamp when data was transferred to the indoor module, "aggregated" is the 5 min aggregation from Netatmo (see Table 2).

Table 2 Example of the Netatmo software that resamples data to regular 5-min intervals, by assigning it to the next full five-minute interval.

Raw	
Time	Rain (mm)
2024-10-29 05:54:48	0.0
2024-10-29 05:59:56	0.303
2024-10-29 06:05:03	1.313
2024-10-29 06:09:58	2.626
2024-10-29 06:15:07	2.626
2024-10-29 06:20:01	1.111
2024-10-29 06:25:09	0.505

Aggregated	
Time	Rain (mm)
2024-10-29 05:55	0.0
2024-10-29 06:00	0.303
2024-10-29 06:05	Not included by Netatmo
2024-10-29 06:10	3.393
2024-10-29 06:15	Not included by Netatmo
2024-10-29 06:20	2.626
2024-10-29 06:25	1.111

2024-10-29 06:30:17	1.616

2024-10-29 06:30	0.505
2024-19-29 06:35	1.616

Between 06:00:00 and 06:05:00 no data is transferred (seen from the column 'raw'). When Netatmo resamples this to a regular time interval, this timestamp is excluded. This means that rain is not actually missing and will not result in an underestimation.

In addition, we suspect that when there is a connection failure between the indoor and outdoor module, the rainfall will be attributed to a timestamp when there is a connection again, potentially aggregating it over a longer time interval than approximately 5 min (see Table 3). However, when the indoor module also has a connection interruption, data is lost.

Table 3 Example of when there was likely a temporary connection interruption between the indoor module and rain gauge. Rainfall will likely be attributed to a timestamp when there is a connection again, resulting in a longer temporal resolution, e.g. at 20:26:21 the rainfall is likely aggregated over 22 min.

Raw	
Time	Rain (mm)
2024-29-10 19:50:59	0.303
2024-29-10 19:56:07	0.101
2024-29-10 20:01:02	0.404
2024-29-10 20:04:08	0.707
2024-29-10 20:26:21	7.777

Aggregated	
Time	Rain (mm)
2024-29-10 19:55	0.303
2024-29-10 20:00	0.101
2024-29-10 20:05	1.111
2024-29-10 20:10	Not included by Netatmo
2024-29-10 20:15	Not included by Netatmo
2024-29-10 20:20	Not included by Netatmo
2024-29-10 20:25	Not included by Netatmo
2024-29-10 20:30	7.777

We suspect that the rainfall attributed to 20:26:21 is actually the rainfall over approximately 22 min.

As we were not aware of this when we wrote the original manuscript, we included these tables (Tables 2 and 3) in our manuscript in the appendix (Tables A1 and A2). We included our findings in Section 2.1, lines 121-128: **"The default rain gauge processing software records**

the number of tips over approximately 5-min intervals, which is communicated wirelessly to an indoor module. Next, the data is transmitted via wifi to the Netatmo platform. The Netatmo software resamples this to regular 5-min intervals, by assigning it to the next full five-minute interval. When within a 5-min interval no data is transferred, this time interval is not included (see supporting information Table A1 for an example). When there is a connection failure between the rain gauge module and indoor module, the rainfall will likely be attributed to a timestamp when there is a connection again, potentially aggregating it over a longer time interval than approximately 5 min (see supporting information Table A2 for an example). However, when the connection of the indoor module is also temporarily interrupted, data is lost."

In summary, the qualification "missing data" is not necessarily the correct term, therefore, we rephrased this. PWSs frequently suffer connection issues and irregular data, using stricter data availability requirements limits the number of PWSs that can be used. In section 4.3, lines 316-320, we rephrased it to:

"A small part of the dataset obtained from Netatmo (5.38% of the selected events' total time steps) were not included, either suggesting that data was missing or that the system suffered connection issues, resulting in irregular data transfer (longer than approximately 5 min). It is expected that the effect of this on the bias is limited, as most of the data is likely not missing, rather caused by irregular data transfer and or connection interruption between the rain gauge and indoor module (see supporting information Tables A1 and A2)."

b) *Did you check whether the PWS used in this study were calibrated by the user, i.e. if the amount of each tip differs from 0.101mm? This might also have an influence on the bias. For consistent results, only PWS with the default calibration could be used - or the number of tips could be determined from the PWS where a calibration was done by the user and converted to the original default value.*

In the metadata there is no explicit information indicating if a PWS is manually calibrated or not. However, we can assume that a PWS is manually calibrated if the tipping bucket volume is different from the default (0.101 mm/tip). To assess the effect of manual calibration, we calculated the number of tips and convert it to the default value. We wrote a small section about the effect of the manual calibration in section 5.1.3. The effect largest on the bias, resulting in a larger overall bias. The coefficient of variation and correlation coefficient remain unchanged or slightly improved (0.01 improvement in CV).

"Around 1/7 of the PWSs used in this study (15%) were manually calibrated by their owner. However, it is unknown what the accuracy of such a manual calibration is. The number of tips was determined for each manually calibrated PWS and converted to the original default value of 0.101 mm. On average, there is a 4% decrease in the observed rainfall by the PWS cluster, resulting in a slightly increased underestimation or slightly decreased overestimation by the PWSs. The CV values slightly improve with an average of 0.01, while the change in r is negligible, see supporting information, Table C1."

c) *Section 5.4. All rain gauges (AWS and PWS) suffer from undercatch alike, maybe PWS a bit more depending on how they are installed. This section could be omitted from my point of view as it offers no new insights.*

This indeed does not provide any new insights. However, it is important to acknowledge that these values from AWSs are not the absolute truth. We moved the text from section 5.4 to section 2.2.

5) Specific comments:

a) *l.27ff consider deleting the brackets or rephrasing in more general terms*

We made this more general in lines 29-31: "**Especially small, fast-responding catchments require accurate rainfall observations with high spatial and temporal resolution for reliable predictions, such as in the order of kilometres and minutes for catchment areas of a few square kilometres."**

b) *l.40ff Radar attenuation should be mentioned here, this is a major source for underestimation of radar data.*

We included this in the introduction, in lines 42-45:
"However, radar rainfall estimates are prone to substantial uncertainty and bias due to several sources of error. These are related to for example the calibration of the instrument itself, signal attenuation and to the conversion from measured reflectivities aloft into rainfall rates at the ground (Uijlenhoet and Berne, 2008; Krajewski et al., 2010; Villarini and Krajewski, 2010)."

c) *l. 47 Are these all PWS? Or just the ones with rain gauges?*
These are indeed only from PWSs with a rain gauge, we made this more explicit by changing it to (lines 48-50): "**The popularity of these low-cost sensors equipped with a rain gauge has been increasing during the last decade, up to around 1 PWS per 9, 11, 13 and 15 km^2 in May 2024, in the Netherlands, Denmark, Switzerland and Germany, respectively.**"

d) *l.62 use*
We changed it to "**use".**

e) *l. 78 El Hachem et al. 2024 is published in the meantime, please update the citation: El Hachem, A., Seidel, J., O'Hara, T., Villalobos Herrera, R., Overeem, A., Uijlenhoet, R., Bárdossy, A., and de Vos, L.: Technical note: A guide to using three open-source quality control algorithms for rainfall data from personal weather stations, Hydrol. Earth Syst. Sci., 28, 4715–4731, https://doi.org/10.5194/hess-28-4715-2024, 2024.*
We updated this citation accordingly.

f) *l. 89 Consider rephrasing this sentence, correcting limitations sounds awkward. Furthermore the the potential of PWS is not only limited to forecasts, it is also a major benefit for hydrological modelling, urban hydrology etc.*
We changed this sentence in the introduction, lines 95-96: "**Quantifying the limitations of PWS rainfall observations and addressing them enhances the potential of PWSs for a wide range of applications, including hydrological modelling, urban hydrology and (hydrological) forecasting.**"

g) *l. 107 maybe mention that is this done via software by changing the volume per tip.*
We included this in the manuscript, see Section 2.1, lines 115-117: "**These gauges can also be calibrated manually by the owner by changing via software the volume per tip, resulting in deviating tipping bucket volumes (approximately 13.5% is manually calibrated according to De Vos et al., 2019).**"

h) *l. 109 consider rephrasing, e.g. "the data is transmitted via WiFi to the Netatmo platform"*
We changed it to (see Section 2.1, line 122): "**Next, the data is transmitted via wifi to the Netatmo platform**."

i) *l. 126f mentioning this limit here is not necessarily required.*
We removed this limitation.

j) *l. 128 „exact recording stations"? You are referring to the PWS that were online at that time I suppose.*
We indeed only have an indication about PWSs that are online at the moment of access, not about PWSs which were for example in operation until 2022. To make this more clear, we rephrased it to (see Section 3.1, lines 144-146): "**Note that the API only provides access to data from PWSs that were operational at the time of access, which was in February 2024. We do not have access to data from stations that were previously in operation but are no longer online at the time of access."**

k) *l. 153ff you are using „selected" quite a few times, consider rephrasing*
We rephrased the sentence (see Section 3.2, lines 172-173): "**The events were selected**

in such way that for the same station and accumulation interval no overlapping time series were included."

l) *l. 172f How many time steps were removed by applying the HI and FZ filter?*
We included the number of timeseries that had at least one HI or FZ flag, see answer 5m).

m) *l. 217f try to avoid 2 -line paragraphs. And consider providing more detailed information on the retained data (with respect to seasons, duration, etc.) here.*
We provided more detailed information about the data after quality control, how much HI and FZ flags were attributed and the occurrence of the manual calibration, see Section 4, lines 270-275**: "After applying the HI and FZ filters and requiring a minimum data availability before aggregation, around 88% of the original dataset was kept. For 87 (0.5%) of the total timeseries used, at least one HI flag was attributed to a timestep. In 93% of the cases, no data was transferred for at least 15 min prior to the flagged HI timestep, suggesting that these flags may result from comparing data aggregated over longer time intervals (≥ 15 min) to a 5-min timestep, potentially leading to mismatches and flagging data. For 5.8% of the timeseries, at least one FZ occurred. Around 15% of the PWSs were manually calibrated, with a median tipping volume of 0.117, with 95% of the calibrated tipping bucket volumes ranging between 0.09 and 0.203."**

n) *Figure 7 It's difficult to distinguish the number of PWS in this plot. I suggest to use a different more differentiable colour ramp*

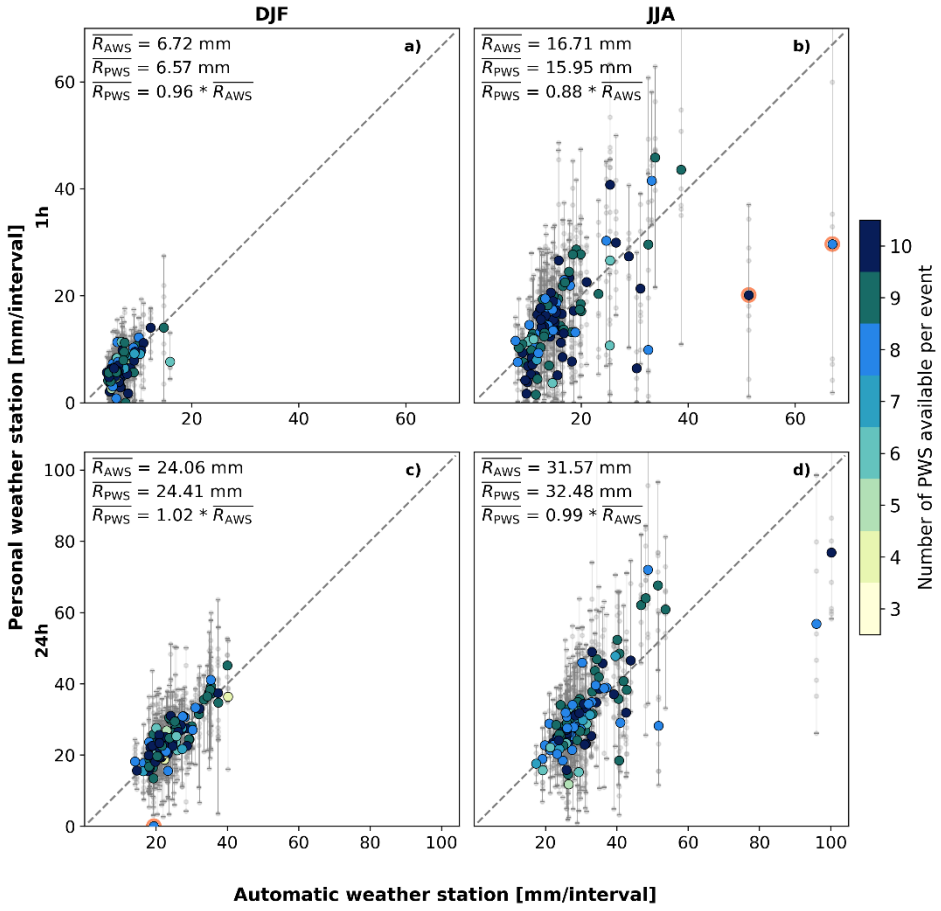

Figure 3 Scatter plots of filtered PWS rainfall accumulations against AWS records for the winter (a, c) and summer (b, d) seasons and accumulation intervals of 1 h (a, b) and 24 h (c, d).

We adjusted the figure (and appendix figures B1 and E1) to make this more clear. See Fig. 3 for the example.

o) *l. 170ff High influx can also result from a connection interruption between the rain gauge module and the base station.*

We are thankful for the reviewer to point this out, as we were not aware of this before. We checked some timeseries and indeed suspect this behaviour. We tested this and noticed similar behaviour. However, only when there is a connection interruption between the rain gauge module and base station, this occurs. When the base station also has a connection interruption, data is lost. We clarified this issue in Section 2.1 (Personal weather stations), see answer 4a.

We checked the number of times a HI was attributed and if the aggregation time was longer than approximately 5 min. In total, in 0.5% of the used PWS timeseries, one or more HI flags were attributed, in 93% of the total cases, no data was transferred for at least 15min or longer. We reported this in Section 4, see comment answer 4m.

p) *l. 184 ff Instead of percentages consider stating how many records per aggregation were required for the valid value (e.g. 10 out of 12 5 minutes for one hour)*

We included both the percentage and the number of valid timesteps that would be required, see lines 245-246.

q) *l. 259 ff Could you please provide more details or discuss why longer transferring and processing errors are reduce for longer intervals or be more specific about intervals for which this holds*

When there is a connection failure between the indoor and outdoor module, the rainfall might be attributed to a timestamp when there is a connection again, aggregating it over a longer time interval than approximately 5 min. For that reason, aggregating over longer time intervals reduces these processing errors. We made this more clear, see Section 4.4, lines 337-341: **“Data transferring and processing errors reduce for longer accumulation intervals, as the effect of attributing rainfall to an erroneous time stamp decreases. This takes place for example when the connection between the indoor and outdoor module is temporarily interrupted, potentially attributing rainfall to a timestamp when there is a connection again, as a consequence aggregating it over a longer time interval than approximately 5 min (see supporting information Table A2).”**