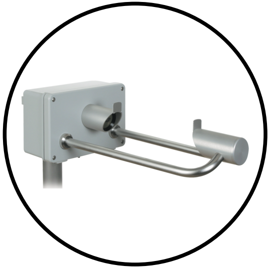
How trustworthy are rain gauge and disdrometer data?

A hierarchical framework for assessing consistency between observations and identifying promising periods for radar calibration

RPG Doppler Cloud Radar Workshop

Delft, June 18, 2026

Dr. Marc Schleiss, Dept. of Geoscience & Remote Sensing, TU Delft



Motivation

Disdrometers and rain gauges are useful references for monitoring the quality and/or calibrating weather and cloud radars. They are pretty reliable overall but **cannot be fully trusted** either!

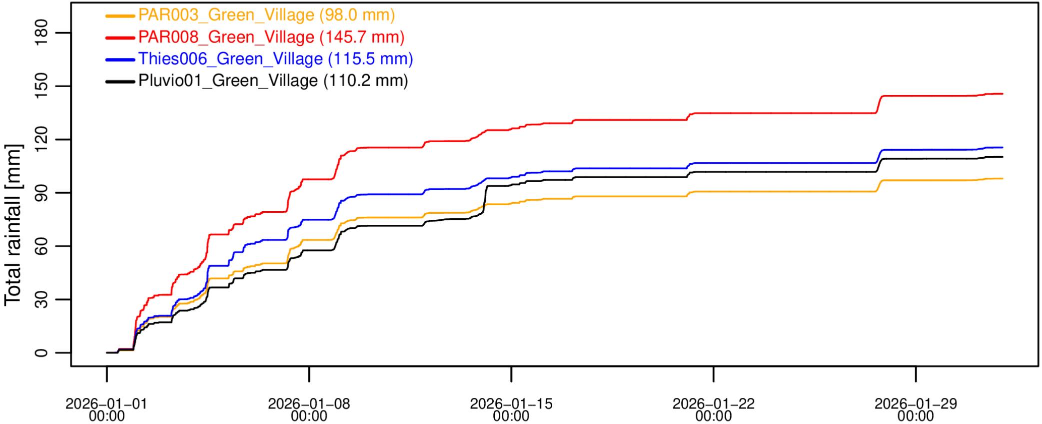
Challenge 1: natural variability

Challenge 2: measurement uncertainty

Challenge 3: different measurement volumes/principles

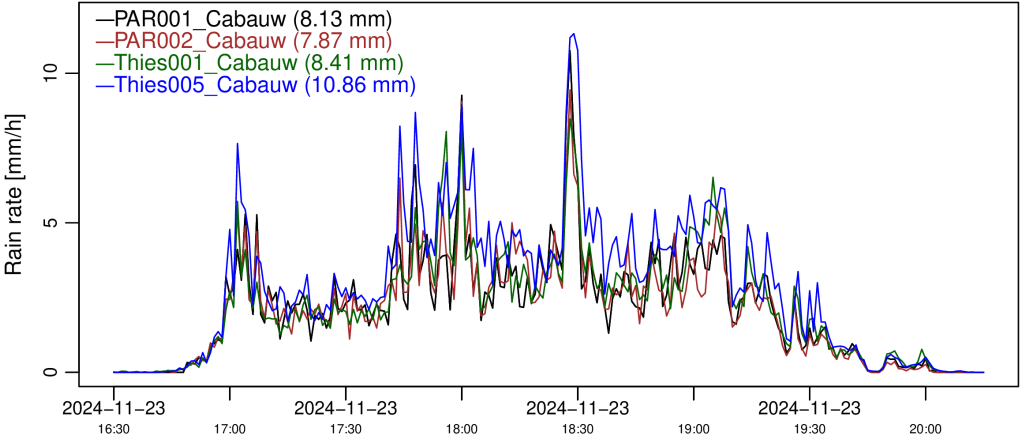
Challenge 4: sensor malfunction/calibration ´

Fact 1: Co-located rainfall sensors do not always agree with each other..



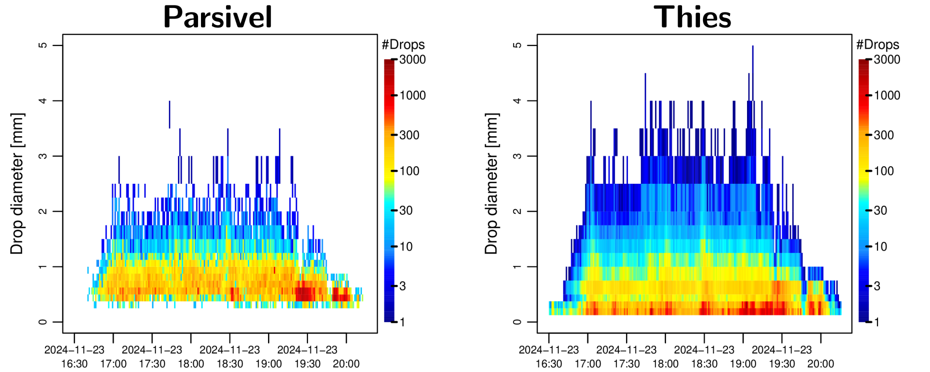
Accumulated precipitation amounts for 2026-01 at Green Village Delft recorded by two OTT Parsivel-2 disdrometers, one Thies Clima disdrometer and one Pluvio weighing rain gauge.

Fact 2: sensor agreement depends strongly on the selected event.



Rain rates measured by two Parsivel and two Thies disdrometers during a stratiform event in November 2024 at Cabauw.

Challenge: even when rain rates agree, substantial differences in drop size distributions (DSD) can occur both within and between disdrometer types. For example, the Thies tends to see a larger spread in DSD compared to the Parsivel.



Average drop size distributions measured by Parsivel and Thies during a stratiform event in November 2024 at Cabauw.

Objective

A key first step in radar calibration is identifying trustworthy periods when disdrometers, gauges, and radar measurements consistently agree with each other.

The goal of this lecture is to introduce some new ideas for how to do this. It is mostly intended to stimulate a discussion rather than provide a definitive answer!

Open Discussion

Imagine we install a rain gauge, a Parsivel, and a Thies disdrometer side-by-side. What evidence would convince you that the data are trustworthy?

Consistency isn't the same as accuracy

When comparing radar, gauges, and disdrometers, the key question is not which sensor is correct, but whether their observations are consistent with one another.

The definition of "consistent" depends on the intended application. It may be in terms of:

- Rainfall totals or intensities, reflectivity, mean drop size, drop size distributions, ..

In the literature, consistency is often assessed through **pairwise** comparisons of sensors (i.e., mutual agreement).

- Bias (i.e., the average difference)
- Correlation (between pairwise values, over a given time window)
- Variability (e.g., standard deviation of pairwise differences)

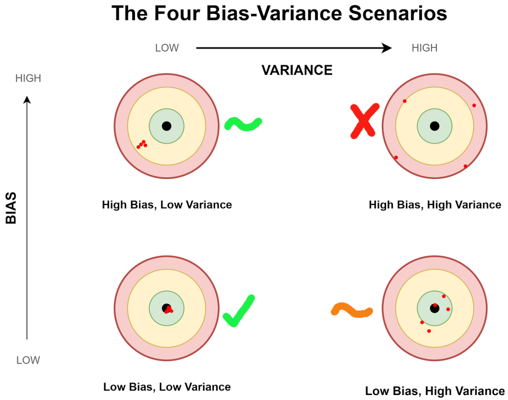
$$\boxed{\text{Consistency} = \text{Co-fluctuation} + \text{small and stable bias}}$$

Small bias vs stable bias

Question: which scenario would you prefer?

A: Two disdrometers that differ by 30%, but whose differences are very stable over time.

B: Two disdrometers that differ by only 1% but whose differences vary by >50%



Answer

Two rainfall sensors with a large stable bias can be more valuable for calibration than two unbiased sensors with large error variance.

Multi-layer quality control system

The following slides introduce a new, hierarchical framework developed by the Ruisdael observatory at TU Delft to assess the consistency of gauge-disdrometer data.

Layer	**Purpose**
L1: Disdrometer-gauge	Estimate random and systematic errors in rain rates and rainfall totals using the rain gauge as reference.
L2: Disdrometer-disdrometer	Assess how repeatable measurements are between disdrometers of the same type (rain rates + DSD-related variables)
L3: Cross-disdrometers	Assess how reproducible measurements are across disdrometers of different types (rain rates + DSD-related variables)

The most trustworthy periods are the ones for which the system is consistent across all 3 layers at the same time.

Layer 1: Disdrometer-Gauge

The rain gauge is the most trusted sensor and therefore acts like an **anchor** (i.e. a baseline against which the system-wide bias and consistency can be compared).

Rain Gauge

$$R_G(t) = R(t) + b_G(t) + \varepsilon_G(t)$$

$R(t)$ = True rain rate at time t
$R_G(t)$ = Measured rain rate
$b_G(t)$ = systematic error (bias)
$\varepsilon_G(t)$ = random error (noise)

Disdrometer

$$R_D(t) = R(t) + b_D(t) + \varepsilon_D(t)$$

$R(t)$ = True rain rate at time t
$R_D(t)$ = Measured rain rate
$b_D(t)$ = systematic error (bias)
$\varepsilon_D(t)$ = random error (noise)

Pairwise differences between disdrometer and gauge

$$\Delta R(t) = R_D(t) - R_G(t) = \underbrace{b_D(t) - b_G(t)}_{\text{systematic}} + \underbrace{\varepsilon_D(t) - \varepsilon_G(t)}_{\text{random}}$$

Performance Metrics for Layer 1

For any given time window, we define the following metrics:

Correlation: $\rho_{DG} = cor(R_D, R_G)$

Bias: $B_{DG} = mean(\Delta R)$

Relative variability: $CV_{DG} = \dfrac{std(\Delta R)}{|B|}$

$R_G(t)$: gauge values
$R_D(t)$: disdrometer values
$\Delta R(t) = R_D(t) - R_G(t)$ pairwise differences

Example: Disdrometer-Gauge Comparisons at Green Village

The following example illustrates Layer-1 quality control for the Ruisdael test site at Green Village

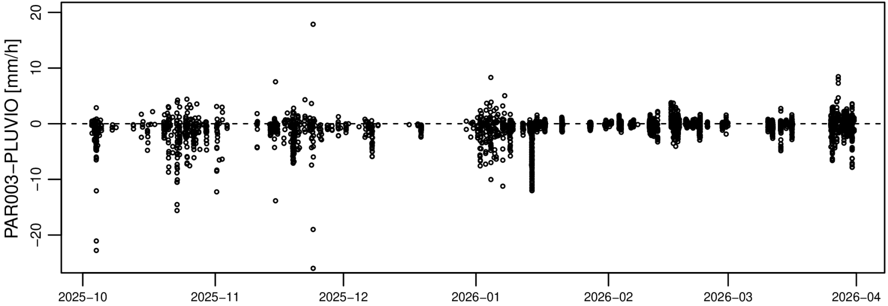
The Experiment

Co-located data from October 2025 to March 2026 from the following sensors:

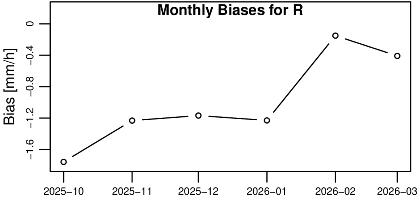
- 2 Parsivel disdrometers (PAR003 & PAR008)
- 1 Thies disdrometer (Thies006)
- 1 Pluvio rain gauge (10-minute resolution)
- $0.1 < R < 100$ mm/h and $10 < Z < 55$ dBZ

Results: Disdrometer-Gauge Comparisons at Green Village

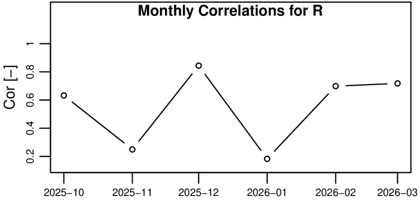
The graph below shows the differences in 10-minute rain rates between the Parsivel disdrometer PAR003 and the Pluvio gauge at the Green Village.



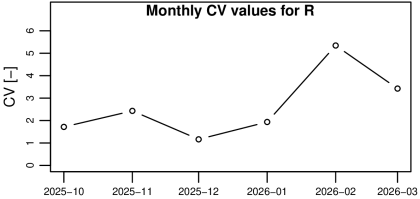
The rain rates do not always agree with each other and the differences fluctuate quite a lot over time. → **Let's compute some Layer 1 metrics!**



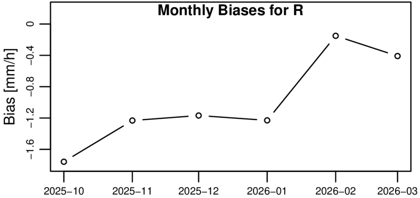
The monthly bias values are negative. PAR003 measures **lower** rainfall amounts than the Pluvio.

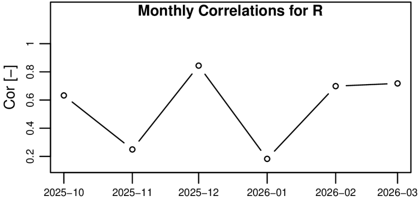


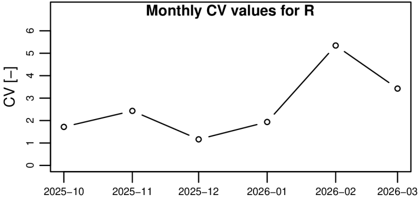
The monthly correlation values vary a lot. November 2025 and January 2026 are the least consistent months.



The relative variability of the differences also varies a lot over time. February and March 2026 have higher CV.







Which is the most trustworthy month according to Layer 1?

Answer: December 2025 has the highest correlation and lowest CV. The bias is quite large but at least the differences are consistent over time.

Layer 2: Disdrometer-Disdrometer

Layer 2 looks at the pairwise differences for two **co-located** disdrometers of the **same type**:

$$\rho^R_{DD} = \mathrm{cor}(R_i, R_j)$$
$$B^R_{DD} = \mathrm{mean}(R_i - R_j)$$
$$CV^R_{DD} = \mathrm{CV}(R_i - R_j)$$

Added value of layer 2

Layer 2 provides information about the consistency of rain rates measured by two identical sensors as well as other DSD-related quantities such as reflectivity Z, mean drop diameter D_m, ..

$$\rho^Z_{DD}, B^Z_{DD}, CV^Z_{DD}$$
$$\rho^{D_m}_{DD}, B^{D_m}_{DD}, CV^{D_m}_{DD}$$
$$.....$$

Main purpose of Layer 2

In addition to providing information about DSD-related quantities, Layer 2 also provides a useful baseline against which Layer 1 (disdrometer–gauge) discrepancies can be interpreted.

Layer	Bias	Correlation	Relative variability
L1 (Disdrometer–Gauge)	B^R_{DG}	ρ^R_{DG}	CV^R_{DG}
L2 (Within-disdrometer)	B^R_{DD}	ρ^R_{DD}	CV^R_{DD}

Interpretation:

- ρ^R_{DD} and CV^R_{DD} provide useful upper/lower bounds for the maximum/minimum possible correlation/scatter between two sensors.
- If $0.9 \leq \dfrac{CV^R_{DD}}{CV^R_{DG}} \leq 1$, then the scatter in layer 1 can be explained almost exclusively by measurement uncertainty.

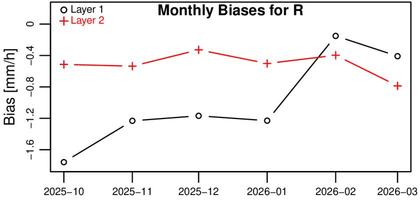
Illustration: Layer 1 & Layer 2 metrics at Green Village

The Experiment

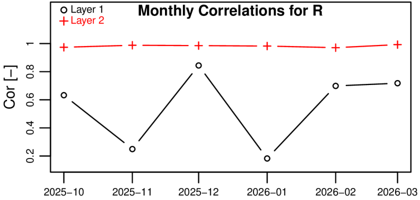
Co-located Parsivel and Pluvio data at Green Village are compared between October 2025 and March 2026.

Layer 1: Disdrometer-gauge
PAR003 vs Pluvio (10-minute resolution)

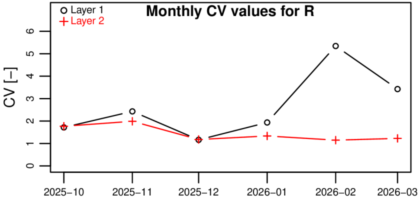
Layer 2: Disdrometer-Disdrometer
PAR003 vs PAR008 (10-minute resolution)



As expected, the average difference in rain rate between two co-located disdrometers (layer 2) is smaller than for layer 1 (Disdrometer-Gauge). ✓



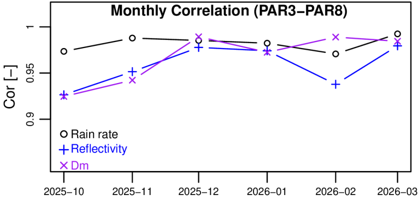
The correlations between two disdrometers of the same type (layer 2) are larger than for layer 1. ✓



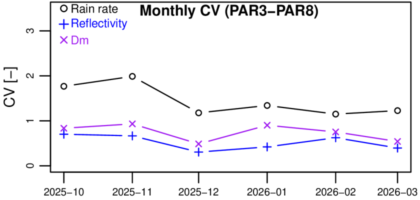
The relative variability of the differences in rain rate in layer 2 is smaller or equal to the variability in layer 1. ✓

Additional benefits of Layer 2

In addition to rain rates, Layer 2 (Disdrometer-Disdrometer) also provides insight into the consistency of other, DSD-related quantities such as reflectivity Z and D_m.



The correlation values for Z and D_m should be above 0.9 and comparable to those for the rain rate. December 2025, January 2026 and March 2026 are the most consistent.



The relative variability of pairwise differences in R, Z and D_m should be small and roughly contant over time. ✓

Layer 3: Cross-disdrometer comparisons

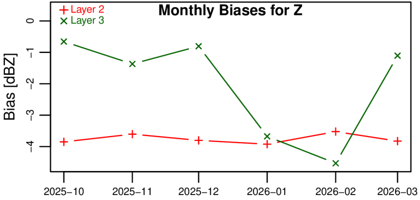
In Layer 3, measurements from two different types of disdrometers (Parsivel-Thies) are compared to each other in order to quantify cross-instrument **reproducibility**.

Terminology

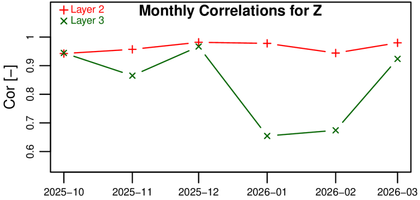
Repeatability: (Layer-2)
Repeatability refers to the consistency of measurements obtained using the same equipment under near identical conditions.

Reproducibility: (Layer-3)
Reproducibility refers to the consistency of measurements obtained under different conditions, such different equipment or environments.

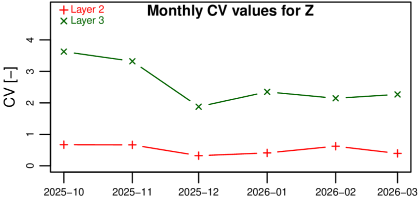
If the discrepancies in layer 3 exceed the within-type variability from layer 2, the DSD measurements are deemed less trustworthy.



The average difference in reflectivity between the Parsivel and the Thies disdrometer (layer 3) is more variable than the average difference between two Parsivel disdrometers in layer 2.



The correlation between reflectivity measurements by the Thies and Parsivel (layer 3) is lower than the correlation between two Parsivel disdrometers (layer 2).



The relative variability of reflectivity differences across disdrometers (layer 3) is larger than the within-disdrometer variability (layer 2).

Putting everything together

In the proposed framework, each layer uses the previous one as a consistency benchmark, creating a hierarchy of trust ranging from **reference agreement** (L1) to **repeatability** (L2) and **reproducibility** (L3). All three layers provide valuable insights. But the key question is: how can we combine them into a single score?

Layer 1	**Layer 2**	**Layer 3**
ρ^R_{DG}	$\rho^R_{DD}, B^R_{DD}, CV^R_{DD}$	$\rho^R_{DX}, B^R_{DX}, CV^R_{DX}$
B^R_{DG}	$\rho^Z_{DD}, B^Z_{DD}, CV^Z_{DD}$	$\rho^Z_{DX}, B^Z_{DX}, CV^Z_{DX}$
CV^R_{DG}	$\rho^{D_m}_{DD}, B^{D_m}_{DD}, CV^{D_m}_{DD}$	$\rho^{D_m}_{DX}, B^{D_m}_{DX}, CV^{D_m}_{DX}$

What to prioritize?

When assessing system-wide consistency, different approaches can be considered:

- A. Agreement between disdrometer and gauge only

  Example of selection criteria: $\rho_{DG}^{R} \geq 0.8$

- B. Agreement + repeatability of rain rates

  Example: $\rho_{DG}^{R} \geq 0.8$ AND $|1 - \frac{CV_{DD}^{R}}{CV_{DG}^{R}}| \leq 0.1$

- C. Agreement + repeatability + reproducibility of rain rates

  Example: $\rho_{DG}^{R} \geq 0.8$ AND $|1 - \frac{CV_{DD}^{R}}{CV_{DG}^{R}}| \leq 0.1$ AND $|1 - \frac{CV_{DD}^{R}}{CV_{DX}^{R}}| \leq 0.1$

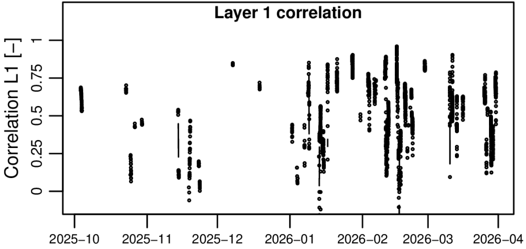
- D. Agreement + repeatability + reproducibility of rain rates AND reflectivities

$\rightarrow$ We can build a decision tree for selecting interesting periods using different criteria and thresholds at each layer

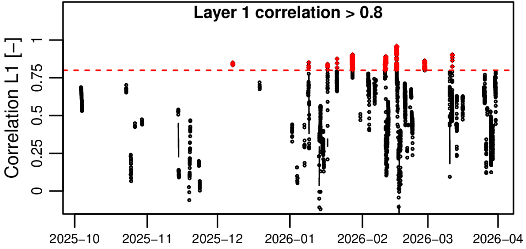
Note: The criteria above exclusively rely on correlation and CV values. The bias is excluded because a stable bias does not make a period untrustworthy.

Example of application

The example below shows the correlation values at layer 1 for the Green Village data.

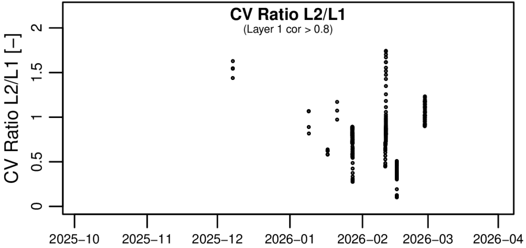


We can select consistent periods by filtering out the periods for which the correlation is below a certain threshold.

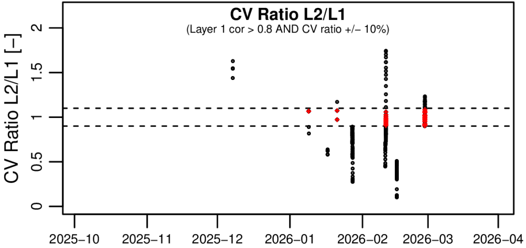


Next, we can refine our selection by looking at the repeatability of measurements!

Selection of trustworthy periods at Green Village

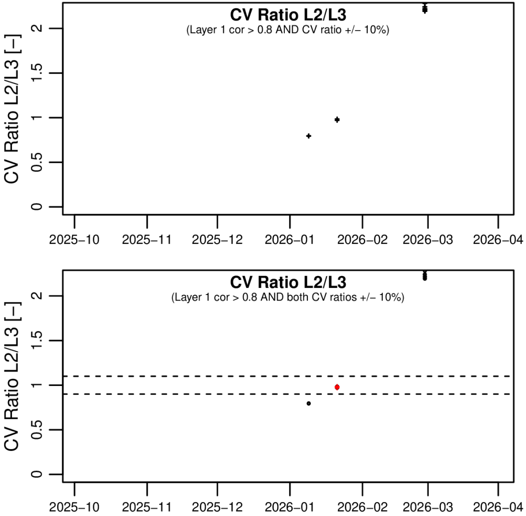


The CV ratio of the rain rate differences in Layer 2 and Layer 1 can be used to identify periods with high repeatability



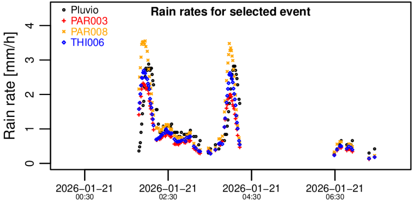
Only the samples for which the correlation and the CV ratios are above/below the thresholds are kept

Final selection

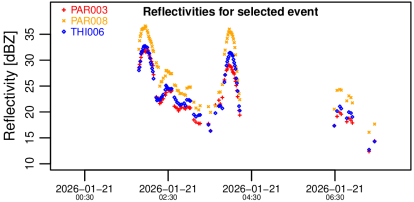


In the final step, the CV ratio of the rain rate differences in Layer 2 and Layer 3 can be used to identify periods with high reproducibility

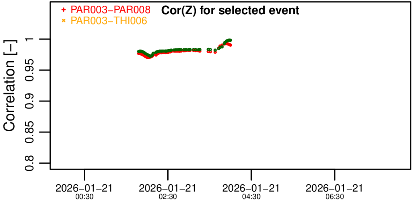
After using all 3 layers, only a single event on 2026-01-21 remains. $\rightarrow$ Let's zoom into it!



The rain rates of PAR003 and the Pluvio are highly correlated. Despite their bias, PAR008 and THI006 can repeat and reproduce the rain rates seen by PAR003.



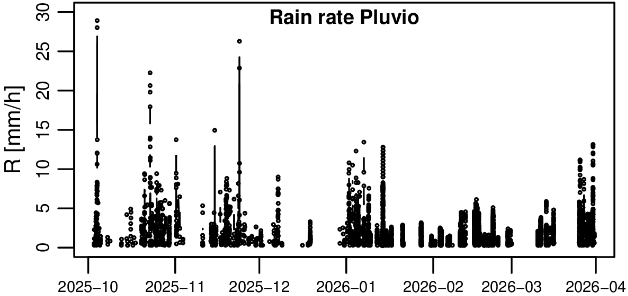
The reflectivity values within/between disdrometers also show a good agreement. Note that we did not use the correlation or CV values of reflectivity for event selection.



Incidentally, the reflectivity values within/between disdrometers also correlate well!

How many trustworthy periods are there?

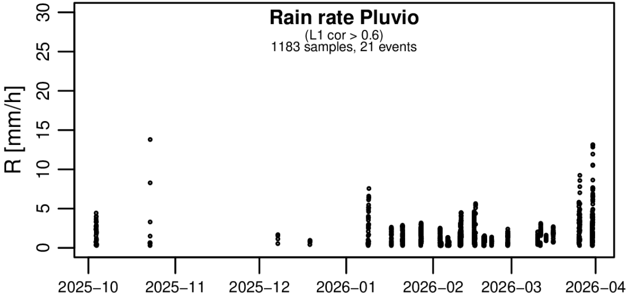
In the previous example, the selection was done using very strict criteria. We can relax these thresholds a bit to find more interesting periods.



Original rain rate time series for the Pluvio, without any filters.

How many trustworthy periods are there?

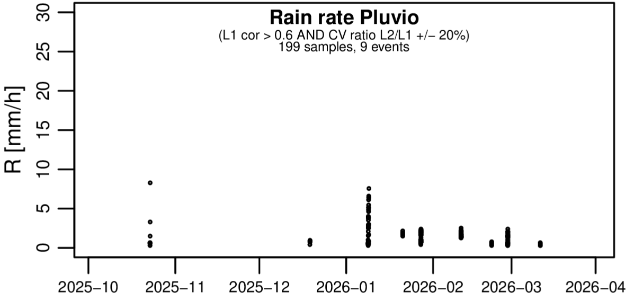
In the previous example, the selection was done using very strict criteria. We can relax these thresholds a bit to find more interesting periods.



After imposing a correlation of at least 0.6 with the gauge, there are 21 events left.

How many trustworthy periods are there?

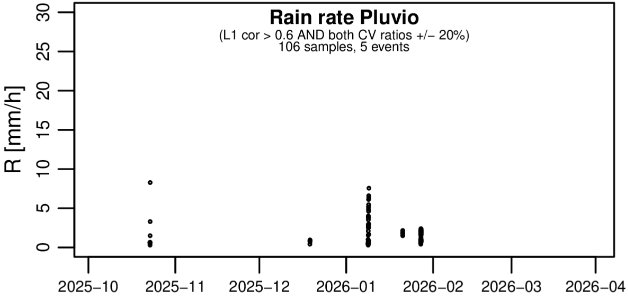
In the previous example, the selection was done using very strict criteria. We can relax these thresholds a bit to find more interesting periods.



After additional filtering on repeatability (20% tolerance on CV), there are 9 events left.

How many trustworthy periods are there?

In the previous example, the selection was done using very strict criteria. We can relax these thresholds a bit to find more interesting periods.



After a final filtering on reproducibility (20% tolerance on CV), there are 5 events left.

By varying the different thresholds and selection criteria, a reasonable number of events can be selected for calibration without the need for arbitrary thresholds on rain rates and reflectivity.

4-layer quality control system?

If available, a 4th layer consisting of reflectivity measurements from a vertically pointing micro-radar could be added to the framework.

Layer	**Purpose**
L1: Disdrometer-gauge	Estimate random and systematic errors in rain rates and rainfall totals using the rain gauge as reference.
L2: Disdrometer-disdrometer	Assess how repeatable measurements are between disdrometers of the same type (rain rates + DSD-related variables)
L3: Cross-disdrometers	Assess how reproducible measurements are across disdrometers of different types (rain rates + DSD-related variables)
L4: Disdrometer-MRR	Assess how reproducible R and Z estimates are when comparing in-situ to remote sensing sensors

Summary and Discussion

Summary:

- A new, 3-layer hierarchical framework for assessing the consistency of co-located gauge-disdrometer data has been introduced.
- The two key metrics are 1) the correlation and 2) the relative variability of pairwise differences of R, Z or D_m.
- The selection criteria can be fine-tuned and adapted to the intended application.
- The method is easy to implement operationally and quite flexible/customizable.
- Don't have 3 disdrometers? No problem! Layers 2-3-4 are optional (but recommended)

What do you think?

- At what temporal resolution should the metrics be computed? Over which time window?
- Should the bias be incorporated into the selection? If yes, how?

Thank you for your attention!

Note: all the data are available here: https://ruisdael.citg.tudelft.nl/

Please send your questions or comments to:

Marc Schleiss (m.a.schleiss@tudelft.nl)