Validating WRF model precipitation phase forecasts using vertical Doppler profilers and disdrometer observations.

Author: Queralt Calderón de Armengol
Supervisor: Joan Bech Rustullet, joan.bech@ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Abstract: In this work we aim to validate data from two model runs of the *Servei Meteorològic de Catalunya*'s WRF model post process using the data of three co-located instruments. These instruments are two Micro Rain Radars (MRR2 and MRR-Pro) and a disdrometer (OTT Parsivel 2). Two *Fuzzy Verification* approaches to resample the data and validate the forecasts were used: one considers the predominant precipitation type during a one-hour interval while the other includes all types which occurred during this period. Contingency Tables were made and Probability of Detection (POD), False Alarm Ratio (FAR) and Gilbert Skill Score (GSS) were used to analyse the model behaviour against lead time and UTC. Results show that the model's performance depends strongly on lead time and on the method used to resample the data. Tendencies were observed comparing both model runs which are discussed considering the limitations of the study.

I. INTRODUCTION

The correct identification of different types of precipitation is important whether for correctly determining the phase of the hydrometeors, taking necessary precautions against adverse weather, or providing input to a hydrological model. When the same variables are obtained by several instruments and models, and using different methods, it is important to determine which one performs better under certain conditions. In this study we aim to compare and validate two model runs of the post-processed product of Weather Research and Forecasting (WRF) model provided by *Servei Meteorològic de Catalunya* using three co-located instruments: two Micro Rain Radars (MRR) and a disdrometer.

The study site is an Eastern-Pyrenees valley, 'La Cerdanya', which is of interest due to its altitude and its topography, which challenges the accurate measurement of solid precipitation using rain gauges. According to (Kochendorfer et al., 2017), solid precipitation can be underestimated using gauges. For this reason, these three instruments were installed in this particular site.

Both MRR datasets were processed using a software named RaProM –see (Garcia-Benadi et al., 2020)– which was developed to obtain the type of precipitation among other variables. Disdrometer data provides the type of precipitation and both model runs also forecast it.

II. DATA USED

A. Study site

The study site is at Das, in 'La Cerdanya' valley. All instruments were located near the Das-Aerodrome weather station –where instruments from previous campaigns have been located–. The Das-Aerodrome weather station is located 1097 m above sea level in the Eastern Pyrenees at the bottom of an inner valley and it is surrounded by mountains that exceed 2000 m. Valleys tend to be relatively isolated from the main air flow thus producing specific phenomena, with thermal inversions being common when the cold air gets caught at the bottom (González et al., 2021).

B. Precipitation types : solid, liquid, mixed and no precipitation

Since three instruments and one model are used for this study, four datasets with different precipitation type classifications are used. For example, the disdrometer can distinguish up to eight precipitation types, whereas the model, just up to four. This is the reason the precipitation was reclassified into three categories corresponding to its phase: liquid, solid, and mixed. In addition to these three, there is the category 'no precipitation'. The program used to process the MRR data also computes the type 'unknown', which was not used.

This phase-oriented classification is not used by any of the instruments or by the WRF model, so a certain criterion was applied to convert data classified using each instrument into our phase-oriented classified data. The procedures followed for each instrument are detailed in the following sections.

C. MRR2 and MRRPRO

Both instruments are Doppler profilers which operate at 24GHz manufactured by the METEK company in Germany (METEK, 2024). They retrieve the vertical profile of Doppler spectra from hydrometeors.

The differences in resolution are:

- For the MRR2, we retrieve 30 altitude intervals of height from 100 m to 3100 m above ground level. These intervals are 100 m high. The time resolution is 1 min.

- For the MRR-Pro, we retrieve 253 intervals of height from 1100m to 7450m above sea level. These intervals are 50 m high. The time resolution is 10 s.

Concerning the measuring frequency of both instruments, the main difference is that the MRR2 is constantly measuring and it provides a daily file in any circumstance, while the MRR-Pro only provides an hourly file when an echo is detected. This process increases the noise in the MRR2 type data. We do not have a file for every hour when the MRR-Pro is measuring, but we do for the MRR2.

D. Disdrometer

The disdrometer is an instrument which uses the principle of extinction to retrieve a spectrum of diameter and fall speed of hydrometeors. It uses the shadow that they cause while crossing a laser beam produced by the instrument. The disdrometer is a Parsivel 2 model provided by HydroMet. This spectrum of diameter and fall speed has 32 classes for both variables. With this data, it is possible to determine the type of precipitation, among other variables. From the classification of precipitation particles, the disdrometer calculates the rain rate. The type of precipitation is based on the number of particles within the measurement range, and the precipitation code is determined from the precipitation intensity R (in mm/h of an equivalent amount of liquid water) (Parsivel Manual, 2025).

The code used for hydrometeor identification is based on the *SYNOP ww Table 4677*. According to this classification there are eight precipitation types: 'Drizzle', 'Drizzle with rain', 'Rain', 'Rain, drizzle with snow', 'Snow', 'Snow grains', 'Soft hail', 'Hail' and 'No precipitation' (Parsivel Manual, 2025).

We considered 'Drizzle', 'Drizzle with Rain' and 'Rain' as 'Liquid'; 'Rain, drizzle with snow' and 'Snow grains' as 'Mixed'; and 'Snow', 'Soft hail' and 'Hail' as 'Solid' for the new phase-based classification.

E. SMC post-process

Catalonia's meteorology agency (*Servei Meteorològic de Catalunya*) runs the WRF model twice a day at 00 UTC and 12 UTC. It is a short-range run since its forecast horizon is 48 h with a time resolution of 1 h. Since it is run twice a day with this forecast horizon we have four simulations per hour (two for each run) except for the beginning and the end of the forecast period.

Each model output is valid for 1 h, it is not a instantaneous forecast. It takes into account the precipitation type and amount, mean temperature and mean relative humidity of the previous hour to compute the next hour ones.

Each model grid cell is defined by its 0.015° separation between points, both in latitude and longitude. The model provides data for all Catalonia but the domain considered in this work is the 3×3 grid closest to Das making the forecast spatially extended data.

The data used from the WRF post process is the type of precipitation. There are four outputs from the model: 'Categorical rain', 'Categorical snow', 'Categorical freezing rain' and 'Sleet' (rain and snow mixed). No cases of freezing rain are forecast the dataset used.

When assigning precipitation types, we assigned Rain as liquid, Snow as solid, and Sleet as mixed.

F. Period of study

The starting date of the period of study is the 21st of November 2023 and the finishing date is the 8th of June 2024 but we only studied the period when all three instruments were active to avoid instrument-biased results.

Frequently, there were hours when some instruments failed to work properly, which were eliminated. For example a long gap occurred from the 8th of March to the 1st of May.

G. Quality Control (QC)

Firstly, there were some points (hourly data) which were eliminated since the data was not reliable.

- For the disdrometer, the points where the detector was not functioning properly (the output of the measurements itself has a parameter indicating the status of the detector and the reliability of the measurement) –e.g., when the anti-freeze heating system failed and the sensor froze–.

- For the WRF post-processed product, the first six hours of each run (corresponding to the first six hours of lead time) were eliminated. The reason for doing so is the model needs time to spin-up. The model starts without any cloud or any hydrometeor species formed and it needs certain time to develop them.

After the invalid data was eliminated there was a QC applied to the data of both MRRs and the disdrometer. The points which did not meet certain conditions were considered noise and their value was changed to 'no precipitation' regardless of its original value. These conditions change depending on the instrument.

- For the disdrometer, all groups of continuous precipitation which lasted less than 6 min were considered noise. This was a QC applied on the dimension 'time' only.

- For both MRRs, the same six-minute interval of continuous precipitation condition is applied and, in addition, the detection of precipitation over a vertical distance of 500 m was required. These conditions were a QC applied over the dimensions of 'time' and 'altitude', since both MRR measures are dependent on them.

Since we are validating the type of precipitation at the surface, we considered the disdrometer to be the reference observation after the QC. The disdrometer was chosen for two main reasons: it is the only one that measures the type of precipitation on the ground, and it is the only instrument which explicitly includes the 'no precipitation' type. For the other instruments 'no precipitation' is extrapolated from the data.

III. METHODOLOGY

A. Fuzzy verification

The main challenges encountered in the data comparison are the difference in time resolution and in position. The disdrometer and both MRRs are co-located punctual measurements while the WRF is based on a grid and their time resolutions are: 10 s (MRRPRO), 1 min (disdrometer and MRR2) and 1 h (WRF). In addition to these instrumentation-related challenges we must also take into account those related to the great spatial and temporal variability of hydrometeors. These data has high variability *per se*.

The method by which the data is sampled can significantly affect the comparison results. In this work we used two *Fuzzy Verification* approaches. In order to *relax* the comparison conditions we downsampled the data in time and space to the lowest resolution of 1 h. It is important to take into account that each instrument has a different time resolution but all of them were downsampled to the lowest, corresponding to the WRF forecasts.

Since we are working with categorical data, we cannot average the variable 'Type of precipitation'. Nevertheless, we applied two alternative approaches to downsample the data. These processes were used to downsample data points in both time (for example, from a resolution of 10 s to a resolution of 1 h) and space when downsampling the WRF post-process from a 3 × 3 1.5 km grid to a single point, as are the other instruments.

1. *First approach: most frequent type of precipitation*

With this approach, only the most frequent type of precipitation was taken into account. The process used is as follows.

- If there were no points with any type of precipitation within the 3×3 grid or during the time interval considered, we assigned 'no precipitation'.
- If a given type of precipitation (A) was repeated more than 70% of the time or grid –when precipitation was detected–, we assigned this type (A).
- If no predominant type was observed, we assigned mixed.

This process is shown in Figure 1 where all the data of one hour becomes one event of liquid precipitation since liquid precipitation is observed 75% of the time.

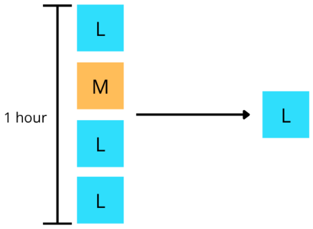

Figure 1. Example of a data resample taking into account only the most frequent type of precipitation. Each colour/letter indicates a different precipitation type: blue (L) liquid and orange (M) mixed.

2. *Second approach: all types of precipitation detected*

With this approach, we consider all the precipitation types detected during the period of time and in the whole 3 × 3 WRF grid. As expected, using this approach we obtained more data points than using the previous one, because when both solid and liquid phases are detected, we obtain two data points for this period of time –one for each phase–. This process is shown in Figure 2, where all the data in one hour becomes three precipitation events.

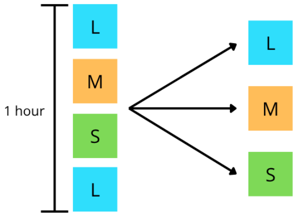

Figure 2. Example of a data resample taking into account all types of precipitation. Each colour/letter indicates a different precipitation type: blue (L) liquid, orange (M) mixed and green (S) solid.

Concerning mixed precipitation, using the previous method, we had two ways of obtaining it: if the most frequent type was actually mixed and if no type was detected more than the 70% of the time. Meanwhile using this approach we only obtained mixed precipitation if this type is explicitly detected.

B. Computed verification scores

In order to validate the models we used three scores: POD, FAR, and GSS whose definition are provided in the Appendix. The formulae and the meaning of each score can be found in (WWRP/WGNE, 2015) and (EUMETRAIN, 2025). Those were computed along contingency tables every time each type of precipitation is validated. The definition for a 2 × 2 contingency table –to validate Yes/No events such as Precipitation/No precipitation– and a 3 × 3 contingency tables –to validate the event of a certain type of precipitation against the other types– are also provided in the Appendix.

POD answers *What fraction of the observed 'yes' events were correctly forecast*, FAR answers *What fraction of the predicted 'yes' events actually did not occur* and GSS answers *How well did the forecast 'yes' events correspond to the observed 'yes' events, accounting for hits due to chance* –which is relevant as it would be easier to forecast precipitation in wetter climates–.

- GSS has a range of -1/3 to 1 where 1 is the best score, but the minimum value depends on the verification sample's climatology. For rare events, the minimum GSS value is near 0, while the absolute minimum is obtained if the event has a climatological frequency of 0.5 and there are no hits. If the score goes below 0 then chance is preferred to the actual forecast, and it is said to be unskilled.

- POD has a range of 0 to 1 where 1 is the best score. This maximum value means there are no misses, so every time the event was observed, it was also forecast. If the value is between 0 and 1 it means the model failed to completely forecast the event. If it is exactly 0, every time the event was observed, the model forecasted something else.

- FAR has a range of 0 to 1 where 0 is the best score meaning no false alarms occurred –a false alarm being the case where the model forecasted the event but it was not observed and there were only hits. If the number is between 0 and 1, the model forecasted the event when it was not observed but it performed correctly on other occasions.

It is important to notice that POD and FAR are complementary scores, meaning that if one is 1 the other is not required to be 0, since they take into account different parts of the contingency table (see Tables V and VI at the Appendix).

Since we are validating a model, the time factor is important. This is the reason we decided to compute the scores against lead time to determine their evolution. Each model run has a maximum lead time of 48 h.

IV. FIRST APPROACH RESULTS: MOST FREQUENT TYPE OF PRECIPITATION

In Sections IV C and V C the results of the GSS versus lead time and UTC are are exposed. When distinguishing between types of precipitation, only the figures where liquid precipitation is validated are shown. The number of data for the other types is much smaller and the fluctuations are larger.

When a shadowed area is plotted in any Figure, this area represents the error variance for each computed validation score.

A. Value counts for the whole period of time before and after Quality Control

Data in Tables I and II shows the number of hours before and after the QC for each type of precipitation. All the instrument data was resampled to a one-hour time period keeping the most frequent type of precipitation. It is important to note that since each WRF run has a 48 h lead time, the number of hours for WRF00 and WRF12, should be twice the instruments' number of hours. If a model run (eg WRF00) forecasts rain at lead time hour 28 (24+4) and, one day after, WRF00 forecasts rain at hour 4 –representing the same UTC hour but a different lead time–,both instances are counted as two separate precipitation events in the following tables.

	Disdrometer	MRR2	MRRPRO	WRF00	WRF12
No preci.	1012	635	650	2682	2590
Preci.	219	1018	1003	240	261
Solid	43	348	733	3	0
Mixed	8	374	185	30	30
Liquid	168	296	85	207	231

Table I. Number of hours for each precipitation type before the QC using the first method of resampling the data.

	Disdrometer	MRR2	MRRPRO	WRF00	WRF12
No preci.	1065	1094	1014	1980	1896
Preci.	156	127	207	179	208
Liquid	115	81	57	147	179
Mixed	7	14	77	29	29
Solid	34	32	73	3	0

Table II. Number of hours for each precipitation type after the QC using the first method of resampling the data.

The comparison of both tables shows that the application of QC significantly reduces the number of hours with precipitation. Since the QC considers noise any interval of continuous precipitation which does not exceed 6 min, it is clear that the noise is substantial, especially in both MRR datasets.

In Table II, after QC, it is shown that the number of precipitation hours is significantly below the one without it, as expected in this climate. The number of hours where liquid precipitation is forecast is quite consistent with the number of times it is observed but this is not the case with the other types. Both runs significantly underestimate the number of hours with solid precipitation, as well as the hours with mixed one with respect to the MRRPRO. In general terms both runs tend to underestimate the hours when precipitation is detected because of the lack of mixed and solid forecasts.

B. Verification scores using the entire time period

To compute the scores in Figure 3, all data was used, after being resampled to a one-hour interval. Hours when all instruments did not detect precipitation were kept since this data does very much have an effect on the value of GSS.

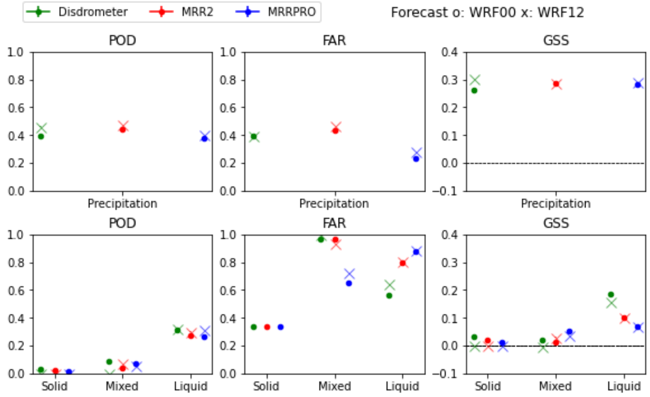

Figure 3. POD, FAR and GSS computed for the whole period of time. In the upper plots we only distinguish between 'precipitation' and 'no precipitation' whereas in the lower plots we divide the 'Precipitation' in three categories.

Concerning POD plots on the left, we can see that roughly 40% of the precipitation events were correctly forecast by both model runs. This number drops to almost zero considering only solid precipitation, but stays near 30% concerning liquid precipitation.

False alarms have a high ratio, especially for liquid and mixed precipitation. Note that there is no false alarm for solid precipitation and WRF12, since this model did not forecast this type, and the value is low for WRF00 since it only forecasted mixed 3 times.

GSS value is slightly better for the disdrometer for liquid precipitation. There is no general tendency for the other cases since all values are really close to zero. From a general point of view, there is no model which is clearly more consistent with any instrument.

C. Study of the dependence of the verification scores on lead time and UTC

Figure 4 shows the evolution of GSS versus the model's lead time separated by instruments. In the left column plots, the validation score is computed while distinguishing between precipitation and the absence of it, while in the right column plot, there is a differentiation by precipitation types inside the 'precipitation' category.

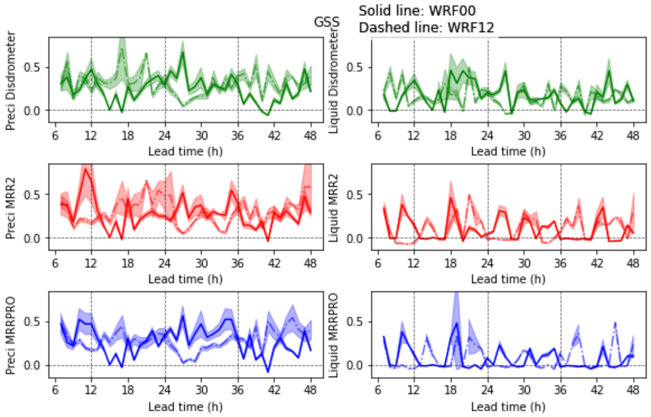

Figure 4. GSS computed for both general precipitation and distinguishing liquid precipitation with respect to the model lead time. Each comparison between the WRF runs and instrument is represented using a distinct colour: green (disdrometer), red (MRR2), blue (MRRPRO).

From a general point of view, we can see that the GSS values for liquid precipitation tend to be lower that the ones for just precipitation, and not distinguishing between types.

Although it is not clearly seen in Figure 4, the same comparison was done plotting instruments against each other (not shown) and the values for the MRRPRO appear to be slightly worse than the other instruments.

In the right plots of Figure 4, the 12 UTC WRF run seems to have a common drop on the GSS around a lead time between 24 and 30 hours for all three instruments. This model run seems to be slightly worse than the other when analysing liquid precipitation, but looking at the right plots in Figure 4, both runs have values really close to zero and the value fluctuates a lot especially for both MRR.

Concerning the error variance S^2_{GSS} for general precipitation, it stays between 0 and 0.2 except for four cases –one for WRF00 and three for WRF12–: WRF00 and the MRR2 at 11 h of lead time where it peaks at about 0.3, WRF12 and MRR2 at 25 h at a value of near 0.3, WFF12 and the disdrometer at 17 UTC peaks at near 0.4, and WRF12 and MRR2 at 48 h of lead time at a value of 0.3.

Regarding the other computed scores, POD appears to be slightly lower for the MRRPRO from a general point of view. POD also appears to have a certain daily pattern when computing it for WRF00 run. When evaluating it for liquid precipitation it rapidly changes from 1 to 0. FAR does not fluctuate as much as POD but it stays high for liquid precipitation, between 0.5 and 1. For general precipitation, FAR fluctuates mostly from 0.75 to 0, depending on lead time.

In Figure 5 is shown the same score but against UTC to examine possible diurnal cycles. The upper plots correspond to the 00 UTC run and the lower ones to 12 UTC one. In the first six hours the 00 UTC seems to slightly

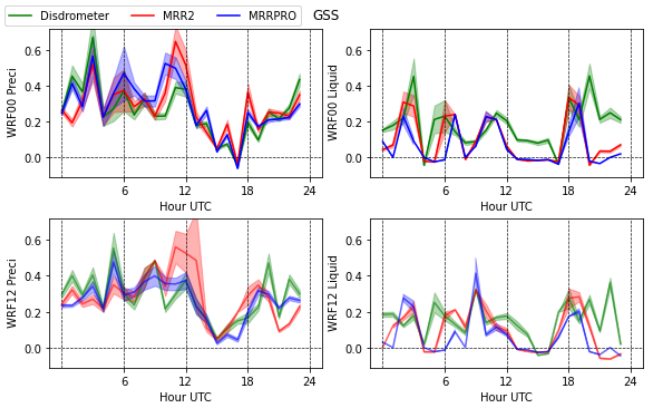

Figure 5. GSS computed for both general precipitation and distinguishing liquid precipitation with respect to UTC.

outperform the 12 UTC one. The value of both model runs drops between 12 UTC and 18 UTC and rises again from 18 UTC to 24 UTC.

D. Precipitation types over the whole time period

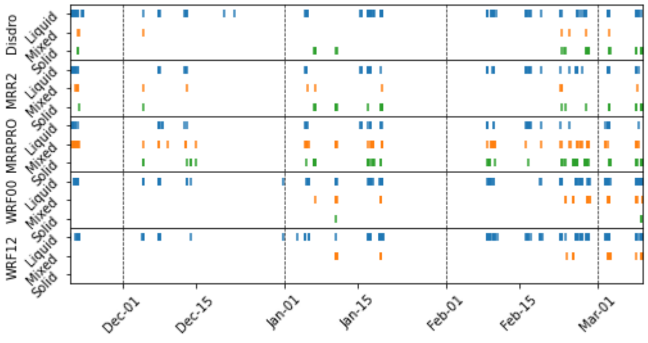

Figure 6. Precipitation types during the whole period for the three instruments and both model runs using the second method of resampling the data. The colours represent different types of precipitation: blue (liquid), orange (mixed), green (solid).

The types of precipitation for the whole period of study were plotted in Figure 6. Each point represents a data point of each type. The lead time is not taken into account in this plot: if two runs of the models forecast different type of precipitation for the same hour –with one model initialized 24 hours after the first one– two data points are plotted.

It is is important to notice that the studied period ends at the beginning of June, but no precipitation was detected by all three instruments apart from the one plotted in Figure 6. Concerning both model runs we can see that solid is barely forecast, while all instruments detect solid precipitation in other episodes of precipitation. The episodes of liquid precipitation are mostly measured by all instruments and forecast by both models.

V. SECOND APPROACH RESULTS: ALL TYPES OF PRECIPITATION DETECTED

A. Value counts for the whole period of time before and after Quality Control

The data in Tables III and IV also shows the number of hours where each type of precipitation was detected after the data was resampled into a one-hour time period but keeping all types when resampling.

	Disdrometer	MRR2	MRRPRO	WRF00	WRF12
No preci.	1014	482	511	1995	1918
Preci.	219	751	722	182	210
Liquid	191	586	303	168	197
Mixed	11	49	231	50	53
Solid	56	597	682	17	20

Table III. Number of hours when each precipitation type is detected before the QC using the second method of resampling the data.

	Disdrometer	MRR2	MRRPRO	WRF00	WRF12
No preci	1068	1092	1016	1987	1906
Preci	159	135	211	182	210
Liquid	137	105	175	168	197
Mixed	10	6	172	50	53
Solid	44	95	200	17	20

Table IV. Value counts for each precipitation type after the QC using the second method of resampling the data.

Comparing Tables III and IV it is clear that, even considering all types of precipitation, the number of hours of precipitation is reduced after the QC for all three instruments but especially for both MRR. The number of hours without precipitation is almost doubled for these instruments.

In Table IV is shown how both model runs continue to underestimate the number of precipitation hours even though the number of hours with each type of precipitation increased using this method of resampling as expected. The number of hours when solid precipitation is detected by the MRRPRO is significantly higher than for any other instrument. The same applies to mixed precipitation. It is important to notice that using this method of resampling the data, mixed precipitation cannot be obtained because no type of precipitation was the most frequent only if mixed was detected.

B. Verification scores based on the entire time period

In Figure 7 we can see that the upper plots of the Figure (those being POD, FAR and GSS for precipita-

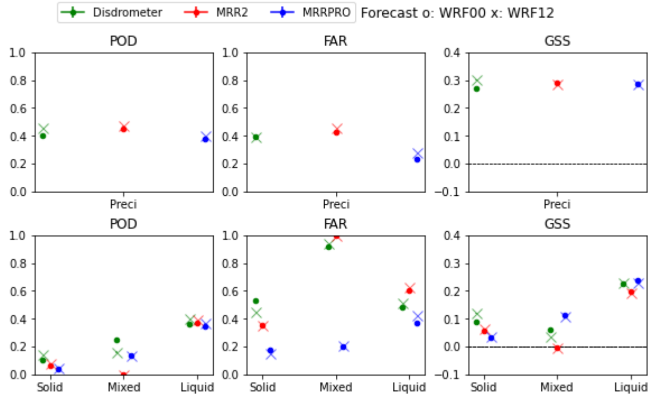

Figure 7. POD, FAR and GSS computed for the whole period of time. In the upper plots we only distinguish between 'precipitation' and 'no precipitation', whereas in the lower plots we divide the 'Precipitation' in three categories.

tion) are the same as when computed using the previous method of resampling. As expected since we also considered the point to be precipitation if there were at least one case of precipitation.

When distinguishing the three different types of precipitation, POD almost drops to zero for solid, and its highest value is for liquid precipitation and WRF12. FAR is quite high especially for mixed precipitation but it drops for the MRRPRO. This is because the MRRPRO detects much more mixed than the others, which can also explain why the GSS mixed value is the highest for the MRRPRO. When liquid is considered, the disdrometer and the MRRPRO have a slightly higher value than the MRR2.

C. Study of the dependence of the verification scores on the lead time and UTC

In Figure 8 the same plot as in Figure 4 is shown. The same differentiation in instruments and in type of precipitation is made.

Looking at all plots in Figure 8, no model run is clearly better than the other for any instrument. It is expected that the accuracy of the model should decrease with respect to lead time but it does not happen, a certain tendency is not clearly observed for precipitation nor liquid precipitation.

The values for liquid precipitation are in a similar range as the ones for just precipitation. This is an improvement with respect the same scores computed for the first approach shown in Figure 4. The value does not fluctuate from zero to another value as much and the same happens for the other type of precipitation (not shown). A similar tendency is shown comparing left and right plots. When WRF00 outperforms when detecting precipitation, the same happens for liquid one.

Concerning the error variance S^2_{GSS}, it stays the same for general precipitation since the same criteria is applied but it slightly increases for liquid precipitation. Despite that, it does not fluctuate as much.

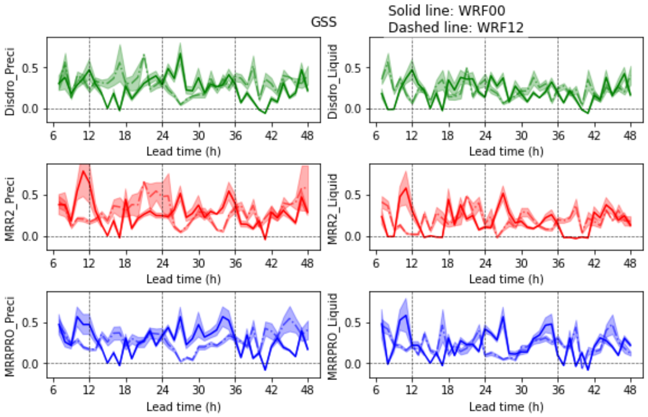

Figure 8. GSS computed for both general precipitation and distinguishing liquid precipitation with respect to the model lead time. Each comparison between the WRF runs and instrument is represented using a distinct colour: green (disdrometer), red (MRR2), blue (MRRPRO)

Regarding other scores computed, the behaviour for POD and FAR are the same for general precipitation but they change evaluating liquid one. POD's daily cycle for liquid was not clearly observed using the first method but now it is more visible, only for WRF00 as in the first method. FAR values are not as high but the fluctuations are bigger.

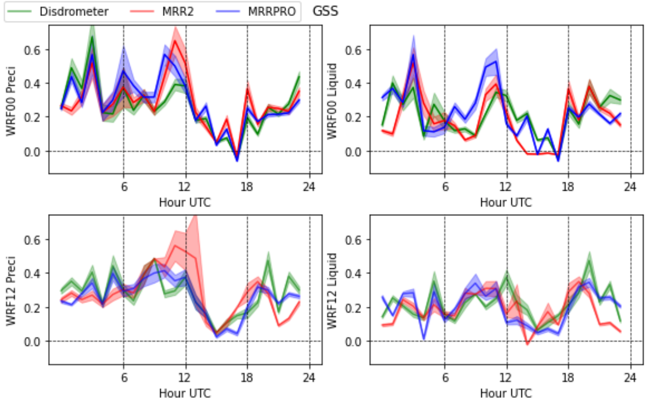

Figure 9. GSS computed for both general precipitation and distinguishing liquid precipitation with respect UTC using the second method to resample the data.

In Figure 9, a similar plot as in Figure 5 is done but the second approach is applied. The left plots are exactly the same (since the criteria is the same as discussed before). The same drop between 12 UTC and 18 UTC is also seen in the right plots corresponding to liquid precipitation. The values of GSS for liquid precipitation are better for the WRF00 run. While values stay quite low, these are better than those obtained using the first method of resampling the data: in this Figure, the behaviour of the right column plots are more similar to the left ones, the fluctuation in the values are not as significant.

D. Precipitation types over the whole time period

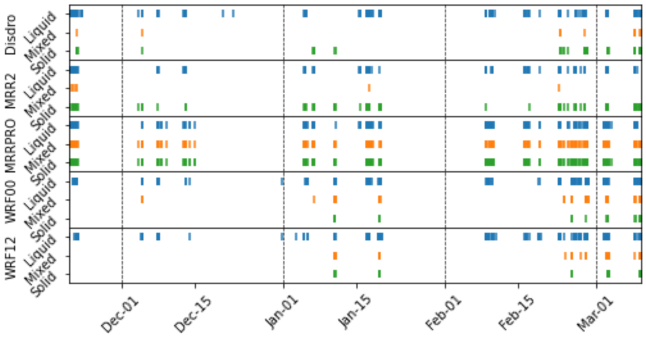

Figure 10. Precipitation types during the whole period for the three instruments and both model runs using the second method of resampling the data. The colours represent different types of precipitation: blue (liquid), orange (mixed) and green (solid).

In general, all instruments detect similar episodes of precipitation but it is clear that MRRPRO sees more precipitation than the other instruments, especially mixed one. That said, zooming into the specific events, it is also seen that mixed precipitation is the least detected by MRRPRO. The two model runs are really similar and they rarely produce a false alarm of precipitation. In Figure 7, we see that FAR for precipitation is 0.4 which is moderately high since this score is computed using hourly comparisons. Meanwhile, in terms of general events of precipitation, the model just has one false alarm at the beginning of February.

In Figure 10 it is shown that the episodes tend to be longer when observing with the MRRPRO. On the other hand, the forecast episodes seem to be shorter than the ones observed.

The WRF runs are more similar using the second approach of resampling the data, since more solid precipitation is forecast. That also makes it more similar to the observation. Both MRRs detect more solid precipitation than the disdrometer and the WRF forecasts.

VI. PATTERNS AND RECURSIVE BEHAVIOUR

During the study, some patterns were observed such as a sudden drop in the GSS value as seen in Figure 5 and Figure 9. A diurnal cycle is seen when computing POD and FAR (not shown), especially when evaluating WRF00. This drop is not as clearly seen when distinguishing liquid precipitation, or any of the other types of precipitation most probably due to the lack of data

Since the phenomenon is independent of the instrument used, it must be related to a factor common to all instruments. It was hypothesized to be due to changes in the number of data available across different hours. We used the data resampled using the first method to investigate the fluctuations in data number. This is shown in Figure 11.

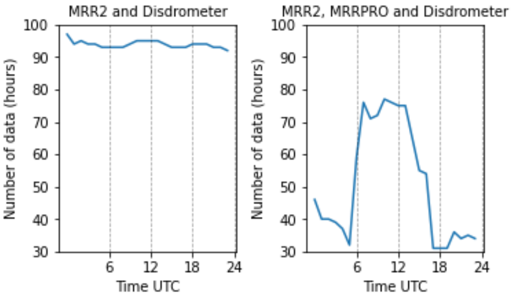

Figure 11. Number of data as a function of UTC hour.

Since it was suspected that there were certain biases in the MRRPRO, we analysed the number of data with and without this instrument. The total number of hours was doubled using only the MRR2, which could simply mean that the MRRPRO was not measuring during a long time period, but when plotting the number of data versus UTC a certain pattern was observed. This pattern is shown in Figure 11. There is a clear peak in the number of data when the MRRPRO is present. To further understand this peak, we obtained the number of hours where each type of precipitation was detected in both situations. There was an increase of at most 5 h in the number of liquid precipitation hours which is negligible against the huge amount of 'no precipitation' data which was added. POD, FAR and GSS were also computed and there were no significant changes.

Since the increase in data is mostly due to noise– specifically, 'no precipitation' data– we eliminated the hours during which no instrument detected precipitation. However POD and FAR presented similar behaviours as before and GSS value was lower (not shown).

VII. CONCLUSIONS

A. Conclusions over the full study period

Generally speaking, the scores for the whole period of time are really similar for both model runs. Observing the global verification scores for both methods in Figures 3 and 7 and the precipitation type detected during the whole period in Figure 6 it is observed that from a general point of view, both model runs can forecast the general episodes of precipitation but fail to forecast solid and mixed precipitation. Mixed seems to be forecasted with a certain time offset. This would explain the high value for FAR using both methods.

Solid precipitation is barely forecast for either models as can be seen in Figure 6, this explains why both POD and FAR are really low for solid precipitation in Figure 3. This value slightly increases when considering all types of precipitation in Figure 7 and it is noticed that the model forecasts more solid precipitation in Figure 10.

The GSS values are considerably better using the second method, which puts an emphasis on how each result is affected by the way we process the data. However, the values are quite low. Liquid is the type with the best value especially for the disdrometer.

When comparing the global value counts using both methods for each type of precipitation in Tables II and IV, –after QC– and also in Figures 6 and 10, when all precipitation types are plotted, it is clear that there is an increase of the number of data for all types, especially for the MRRPRO. This is expected since the MRRPRO has a time resolution of 10 s, so it is easier for this instrument to observe small changes. That likely causes an overestimation of the number of detections, this instrument is too sensitive to apply the second method of resampling the data, at least, with a resampling interval of 1 h.

The fact that the episodes of precipitation seem longer when seen by the MRRPRO in both Figures 6 and 10 –but specially in the second one– can also be explained by this instrument seeing smaller events. If one zooms in to the specific episodes of precipitation, the MRR2 barely sees the smallest events and they can easily be interpreted as noise, whereas the MRRPRO sees those events better defined.

Less mixed precipitation was expected using the second resampling method, as it relies on direct detection rather than the absence of a dominant type. However, this was not observed, likely because this effect was compensated by considering all types in one hour. The number of mixed precipitation is only lower using the second approach for the MRR2.

The fact that both MRRs observe more solid precipitation than both disdrometers and WRF is as expected since the MRR do not detect at the surface but at a certain altitude whereas both the disdrometer and the WRF model provide data at surface level.

B. Conclusions of the study against lead time and UTC

In Figures 4 and 8 we expected the GSS value to drop with lead time but this tendency is not seen. It is also noticed that the values for these scores are generally higher than the global value considering the whole period of time. This two facts can be attributed to the limited number of data. The highest number of precipitation hours is around 200 at most, meaning that, on average there should be around 5 h of precipitation for each lead time hour.

Concerning the GSS fluctuations in Figure 4, these are not as large when we used more data in the second method. This also explains why there is a similar tendency between general precipitation and liquid one in Figure 8, meaning that when WRF00 outperforms WRF12 (and *vice-versa*) when detecting precipitation, the same happens for liquid one.

The error variance S^2_{GSS} is slightly higher using the second approach but it fluctuates less. Regarding the behaviour of other scores, POD cycle is instrument independent, so it can be related to the first run of the model. The change in FAR values is highly related to the number of data available. Using the first method, the temporal offset between observation and forecast probably dominates, which may explain the elevated FAR values. Using the second method, the fluctuations stay important, as the increased detection of liquid precipitation does not fully compensate the offset.

When analysing the behaviour against UTC hour, there is a drop in the GSS value for all runs using both methods of resampling the data. To explain this phenomena we looked into the lead time plots in Figures 4 and 8. This UTC time period when the drop is found corresponds to the following lead times: 12 h to 18 h and 36 h to 42 h for WRF00, 0 h to 6 h and 24 h to 30 h for WRF12. WRF00 is performing worse than WRF12 between 13 h and 17 h and between 36 h and 46 h, meaning that the time where this drop is corresponds to the lead time when WRF00 is performing at its worst.

Looking at the lead times which corresponds from 12 UTC to 18 UTC in WRF12 lead time, the data for the first 6 h are eliminated due to the spin-up process on the model and WRF12 performs worse than WRF00 from 24 h to 30 h. During this interval of time both runs perform worse or the data is eliminated.

C. Conclusions of the study on observed patterns

In Figure 11 it is clear that the MRRPRO has a certain diurnal cycle. This behaviour can be attributed to the fact that MRRPRO only provides an output when an echo is detected, since this instrument is near the aerodrome, it is likely to be influenced by its activity, which should produce more noise during the day. This 'no precipitation' is obtained from the noise of the aerodrome.

MRRPRO's behaviour is leading the trend of the global number of data. However, since the number of hours where precipitation is detected does not change significantly, the patterns should not be explained by the differences in the number of data. Furthermore, this number of data peak is located between 6 UTC and 18UTC whereas the drop in the scores is between 12 UTC and 18 UTC. This drop corresponds to the first 6 hours of modelling for the 12 UTC WRF run and these hours were eliminated. This patterns are likely associated with the model.

VIII. APPENDIX

The formulae for the verification scores used in this work are the following:

$$\mathrm{POD} = \frac{\mathrm{Hits}}{\mathrm{Hits} + \mathrm{Misses}}, \tag{1}$$

$$\mathrm{FAR} = \frac{\mathrm{False\ Alarms}}{\mathrm{Hits} + \mathrm{False\ Alarms}}, \tag{2}$$

$$\mathrm{CSI} = \frac{\mathrm{Hits}}{\mathrm{Hits} + \mathrm{False\ Alarms} + \mathrm{Miss}}, \tag{3}$$

$$\mathrm{GSS} = \frac{\mathrm{Hits} - \mathrm{Hits}_r}{\mathrm{Hits} + \mathrm{False\ Alarms} + \mathrm{Misses} - \mathrm{Hits}_r}, \tag{4}$$

where

$$\mathrm{Hit}_r = \frac{(\mathrm{Hits} + \mathrm{False\ alarms})(\mathrm{Hits} + \mathrm{Misses})}{\mathrm{Hits} + \mathrm{False\ alarms} + \mathrm{Miss} + \mathrm{Correct\ Negatives}}, \tag{5}$$

is the number of hits for a random forecast.

The definitions of a 2 × 2 and 3 × 3 contingency tables are shown in Tables V and VI respectively.

Table V. Standard definition of a 2 × 2 contingency table.

	Event observed	
Event forecast	Yes	No
Yes	Hit	False Alarm
No	Miss	Correct negative

Table VI. Standard definition of a 3 × 3 contingency table to validate event ‘A’ where FA means ‘False Alarm’ and CN means ‘Correct Negative’

	Event observed		
Event forecast	A	B	C
A	Hit	FA	FA
B	Miss	CN	CN
C	Miss	CN	CN

In Table VI the definitions for a contingency table for event ‘A’ validation are shown. To fully interpret it, is important to take into account that this kind of tables are created in a ‘Hit or miss’ context. When validating event ‘A’, if the observation is B (observation miss) and the forecast is also B (forecast miss) the result is a correct negative, since neither the observed nor forecast was ‘A’. When validating event B, the numbers of the contingency table do not change but role they play when calculating the scores does.

ACKNOWLEDGMENTS

Firstly I would like to thank my supervisor Joan Bech, who helped me from the beginning, since I stated working in the PANDORA project. I would also like to thank everyone who was involved in this same project and the people in the Meteorology group at UB who were always willing help with anything I needed. I would like to acknowledge everyone at SMC for providing us with the WRF data and specially J. Mercader who offered his help when using it. And finally, I must acknowledge my friends and family who bore with me while I was complaining about all the problems I had with the plots.

REFERENCES

[EUMETRAIN, 2025] Critical Success Index (CSI) or Threat Score (TS), and Equitable Threat Score (ETS). (n.d.). EUMETRAIN. https://resources.eumetrain.org/data/4/451/english/msg/ver_categ_forec/uos2/uos2_ko4.htm (Last accessed: May 2025)

[Garcia-Benadi et al., 2020] Garcia-Benadi, A., Bech, J., Gonzalez, S., Udina, M., Codina, B., & Georgis, J.-F. (2020). Precipitation Type Classification of Micro Rain Radar Data Using an Improved Doppler Spectral Processing Methodology. *Remote Sensing, 12*(24), 4113. https://doi.org/10.3390/rs12244113

[González et al., 2021] González, S., Bech, J., Garcia-Benadi, A., Udina, M., Codina, B., Trapero, L., et al. (2021). Vertical structure and microphysical observations of winter precipitation in an inner valley during the Cerdanya-2017 field campaign. *Atmospheric Research, 264*, 105826. https://doi.org/10.1016/j.atmosres.2021.105826

[Kochendorfer et al., 2017] Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., et al. (2017). Analysis of single-Alter-shielded and unshielded measurements of mixed and solid precipitation from WMO-SPICE. *HESS, 21*(7), 3525–3542. https://doi.org/10.5194/hess-21-3525-2017

[METEK, 2024] Micro Rain RADAR MRR | MRR-2. (2024, July, 8). METEK. https://metek.de/product/mrr-2/ (Last accessed: May 2025)

[Parsivel Manual, 2025] OTT Hydromet GmbH. (n.d.) Operating Instructions Parsivel Application Software ASDO, [Intruction Manual].

[WWRP/WGNE, 2015] WWRP/WGNE Joint Working Group on Forecast Verification Research. (2015, January 26). Verification. Collaboration for Australian Weather and Climate Research. https://www.cawcr.gov.au/projects/verification/ (Last accessed: May 2025)