Reviewer #3

General comments

The paper was well written, containing a good introduction to readers also outside meteorological and hydrological community. Also the topic – uncertainties in predicting intensive flooding – is very relevant in modern society still sensitive to weather conditions. The geographical scope of this study was limited to Poland, but it can be expected that the results are applicable in Southern and Northern climates as well, especially in mountainous areas.

Specific comments

The experiments section involved verification of products that depend on inputs that are used as ground truth in comparisons. This was clearly pointed out throughout (L377, L426, L485, L548, L556, L582, L621) the article which is of course appreciated. Nevertheless, interpreting verification is problematic in these cases. On one hand, for example, one could easily think of multi-input algorithm design where output if asymptotically forced to match point measurements used also as reference – yielding zero error. Or if the values do vary, it would be good to know motivations in design (for example, avoiding overfitting). Nevertheless, such evaluations have little or no information in my opinion. On the other hand, experts in this area do know the challenge of evaluating input sources (measurement technologies), none of which is perfect (L95) . It is operationally tempting – if not inevitable – to combine inputs of various type (L123). But as to verification of performance, one should try to use reference data as independent as possible. Would it be easy to use some kind of cross-validation, dropping some ground observations (in turn) from input, and to verify the results against those? This could yet require more computational effort.

> We thank the Reviewer for these comments! Of course, we discussed a lot in the authors' team about this topic, and finally:
>
> – We agree that "one should try to use reference data as independent as possible". Therefore, if we use data that is not completely independent, we always include the annotation "(dependent)" in the results tables.
> – We have removed the GRS Clim fields from the verification of the 24-h accumulation because the final step in the development of the product is to adjust it to the GAU Manual, so later verification with the same data cannot give fully reliable results. All other data after removing GRS Clim, are independent of the GAU Manual data.
> – Because it is impossible to use the daily GAU Manual accumulations to verify 1-h accumulations, we chose the 1-h multi-source GRS estimates as reference data, which, as shown in Table 2 shows, are the best estimators of 24-h totals.
> – Thus, in 1-h verifications with GRS as reference we have focused on the reliability of other estimates on which GRS does not depend in any way, e.g. satellite precipitation, from NWP models, and unconventional techniques.

When evaluating prediction based on commercial microwave links (CML), a natural explanation (L423,L584-585) of deviations is distance from reference measurements (GAU manual). Could it be useful to study the effect of distance by measuring correlation inside reference data itself? Then input data from a separate system (like CML) could be then compared against such modelled, "theoretical" maximum – providing estimates of measurement uncertainty at least in the vicinity of the links. (This is more a suggestion for further work, not for this study and this could be of more interest for CML application developers.)

Thank you very much for these suggestions! We are currently working hard on the calibration, quality control, and operational implementation of CML data, which proves to be not so easy. We will definitely use this suggestion of autocorrelation analysis of GAU Manual data.

The article reports errors in predictions using input from radar and especially, satellites. Many potential error sources (L58, L59; L69) are well-known – like measurement geometry or uncertainly of water phase (in both radar and satellite measurements). It would be interested to read the author's views on which of the error sources have been critical in this study. Systematic analysis could be certainly outside the scope of this article, but perhaps just visual inspection could be used as a basis for discussion on error sources.

We have supplemented section 4.2 with our comments on this topic:

to line 416:

"The radar network in the analysed flood area is relatively dense, but due to signal blocking by mountains, precipitation shadows appear in some places, which result in an underestimation of precipitation. This is particularly evident in the Kłodzko Valley which is surrounded by relatively high mountains and is one of the places most prone to catastrophic flooding."

to line 420:

„The reliability of precipitation estimates based on satellite data is low, especially when they are generated from infrared channel data and are not supported by other, preferably microwave data (from radars). This mainly affects SAT estimates, but also others. It should be noted that during the analysed flood, data from visible channels was only available for about 1/3 of the time, due to the fact that for the measurements to be reliable, the sun must be sufficiently high above the horizon (above 20 degrees). Furthermore, the spatial resolution of these data is generally insufficient."

Focusing separately in cases of intensive rainfall (Sec 4.4.) is well motivated. When thresholding the cases with reference (L505, L531), negative bias is reported for all the methods (Table 4), also highlighting it for radar in text (L510, L546). Especially in verifying GAU against GAU Manual, I think that the reported underestimation (L573, L574) is a direct consequence of the applied thresholding! Consider two measurement devices of similar climatology some kilometres apart and long-term statistics of (convective) rainfall: measured rainfall is then similarly distributed over the mean value. But if studied cases are limited by thresholding data on ONE measurement location/device, the other still includes also the lower values, pushing its bias down! (Consider throwing two dice, comparing averages of each, but limiting the studied cases by thresholding the first die.) I guess also with radar, similar effect can be observed when limiting cases by thresholding the reference value. (Radar's bias could be basically zero, but "random noise" ie. positive and negative deviation around mean is now caused by non-uniform vertical profiles of precipitation and advection, for example.) If you agree with me in this, I suggest you somehow address and elaborate this in text and/or presented experiments.

We are aware of the limitations of a methodology of excluding precipitation accumulations that are below a certain threshold from the statistical analysis. We agree that its application requires commentary and an indication of the consequences of these limitations. Thus, in place of the sentence on lines 506-507 in Section 4.4, we have inserted the following fragment:

"As expected, the results are noticeably worse when compared to those obtained without a limitation on precipitation magnitude (see Table 2). This is particularly evident in terms of bias, which indicates an increase in underestimation. However, a negative bias was observed for all the estimation techniques analysed, even without thresholding. This suggests a real underestimation of intense precipitation by these methods, rather than simply a result of data selection."

Technical comments

RainGRS is mentioned several times before explained or referenced. It is also unclear, what is "RainGRS (GRS)" compared to plain "RainGRS".

> We have reordered this. In the abstract where the name “RainGRS” first appeared we have changed the sentence in lines 18-21 to:
>
> > “Reference data used to verify the reliability of the different techniques for measurement and estimation of precipitation included observations from manual rain gauges and multi-source estimates from the RainGRS system developed at IMGW for daily and hourly accumulations, respectively.”
>
> “GRS” are multi-source estimates from RainGRS system – see Table 1.

Long URLs embedded in the text (L328, L337, L345, L353) reduce readability a bit (). If the publisher's style guide supports it, could they be in the references?

> We have included links to references:
>
> line 328:
> NASA. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V07 (GPM_3IMERGHH): https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_07/summary?keywords=%22IMERG%20final%22, last access: 29 July 2025.
> line 337:
> CHRS, https://chrsdata.eng.uci.edu/, last access: 29 July 2025.
> line 345:
> ECMWF. ERA5 hourly data on single levels from 1940 to present: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels, last access: 29 July 2025.
> line 349:
> NCAR. Weather Research and Forecasting model WRF: https://ncar.ucar.edu/what-we-offer/models/weather-research-and-forecasting-model-wrf, last access: 29 July 2025.
> line 353:
> DWD. ICON (Icosahedral Nonhydrostatic) Model: https://www.dwd.de/EN/research/weatherforecasting/num_modelling/01_num_weather_prediction_modells/icon_description.html, last access: 29 July 2025.

A minor detail: place names seem to have mixed style; English names should be preferred if they exist. (According to Wikipedia, Oder seems to be the established English name for the river Odra (PL). Also Sudetes (EN) and Sudety (PL) appear, but understandably the smaller the locations/regions, less English names exist! Anyway, I leave it to the authors to decide the naming policy.)

> We have standardised the nomenclature throughout the article. We have left geographical names in Polish, as they are more commonly used in the literature.