Review of "Accelerating research through community open source software for a standardized file format to improve process representation in numerical weather prediction models"

General comments

The manuscript showcases a toolkit to standardize observational data in a way that would make it easier for model physics developers to utilize observational data from non-standard sources in their work. The manuscript presents two open source Python packages, one dealing with creation of the standardized data ("Merged Data Format") and one with visualization of this data.

The overarching goal presented in the manuscript is commendable – it would indeed be very useful for model developers to be able to verify their model forecasts more easily against different observational datasets such as those made during YOPP as discussed in the manuscript. However, I feel the manuscript in its current state is nowhere near publishable: it currently reads to me more like an internal project documentation rather than a scientific article presenting an innovative new observation handling toolkit (that it should do).

Major issues

1) What is lacking in my opinion is a structure that would allow the reader to understand the logic and actual benefits of using the MDF format. One way of improving on this would be to showcase in detail what the code does through presenting a clear example of the workflow. What I mean by this is to include a detailed example of what the MDF toolkit actually does, an example:

i. Present some observational datasets that do not follow the same (meta) data format (something similar to ncdump -h output). It would be nice to see how a) two different similar observations (e.g. tower data) are harmonized by MDF, b) and how dealing with e.g. sounding data differs to that of a point observation data.
ii. Present in detail how MDF toolkit logic handles reformatting these different datasets.
iii. Present how the (meta) data format looks like after MDF toolkit handling.
iv. Repeat i)-iii) for 1-2 model forecasts.
v. Present how these observations and forecasts in MDF format would be visualized with the MDF visualization tool (show examples of code call and plots generated).

2) netCDF is referenced as the basis for the MDF format, but the manuscript does not state whether the MDF format is actually netCDF with standardized meta data, coordinates etc., or a completely different data format that won't be readable through netCDF libraries and applications (e.g. ncview, GRIB-API/ecCodes). Please clarify.

3) I do not know if this is an issue caused by the language used or do the authors actually share this line of thought, but currently, especially the introduction, presents **all** observations to be problematic for modellers. There should be a clear distinction between observational campaigns and non-operational observational datasets that are (or might be) problematic, and long-running operationally handled observations that already conform to a standard (operational soundings, SYNOP, METAR, …) and are used daily by operational NWP models. Please make this distinction clear.

3) The grammar of the manuscript needs considerable work (punctuation rules etc.).

Minor issue

1) Could the authors expand on why comparing multiple models would be an issue of (P2 L46)?

2) Again, please check your language, constructs like “process based NWP” should not be in (if a NWP model does not include physical parametrizations, i.e. “processes”, they are dynamical models).

3) Table 1: is observational error included and is plotting it a part of the visualization toolbox?

4) Why are you only testing 3 out of your 7 test-cases against standards? Shouldn’t you aim for all pieces of your program to meet these standards?