Review of “Merged Observatory Data Files (MODFs): An Integrated Observational Data Product Supporting Process-Oriented Investigations and Diagnostics” by Taneil Uttal et al.

The MODF concept is laudable and should be supported. The paper as drafted does a good job of discussing the technical details of what an MODF file is. It does, however, skirt over or around a number of fairly hefty issues which it would be worth trying to address in redrafting in my view while, of course, trying not to detract from the main technical nature of the piece.

Major comments

1. The preparation of files in a highly usable format is a necessary but not a sufficient condition for the broad scale use of observatory and campaign data. While absolutely not the focus of the paper it feels very remiss not to have a brief 1-2 paragraph discussion around the broader aspects surrounding exchange, archival and dissemination of these data. Basically, if every single observatory and campaign retrospectively reformatted everything to the proposed format we would still not have solved the broader problems around discoverability and accessibility of these data which requires a systemic effort to collate and provide more unified access to the data. I would think a couple of paragraphs discussing next steps to enable exploitation using the MODF concept as a way to harmonise data formatting issues would strengthen rather than distract from the piece. You’d basically be making the case that MODF is an enabling step in a broader activity to enable greater use of these data by the community. This may include a dedicated effort to collate such data from multiple existing and planned observatories and campaigns and provide access via a single unified repository which may well be federated in a similar manner to CMIP itself.
2. While issues of cadence or reporting frequency are dealt with, it is unclear how broader collocation issues are dealt with in the proposed MODF file format. Take an example of measuring upper-air variables with a lidar, a radiosonde, an FTIR and a monumented GNSS sensor. While all may nominally sense water vapour the measured volumes as well as the time intervals differ substantively (balloons drift, lidars measure vertically, FTIR is in direction of sun. GNSS depends upon multiple complex path angles that are ever varying). It is not sufficiently clear how these distinctions are dealt with in a file or how a user is guided to account for the fact that there will be differences arising from what was measured rather than the measurements themselves. See Immler et al 2005 for further discussion in the context of metrological comparison closures in developing GRUAN products.
3. If MODF files enable version replacement for subcomponents then how are MODF files themselves proposed to be versioned and archived to enable reproducibility? If very actively curated there could be tens or even hundreds of unique versions of MODF files as different subcomponents are periodically reprocessed and reissued? The description is a little unclear to me as given how this will be handled. Maybe its covered in Section 4 but if so its not sufficiently clear to me as presently drafted and it would be beneficial to redraft for clarity.

4. Is the H-K schema a subset of GeoJSON or other emerging standards? It might be worth being a little more explicit. At least some of the names appear to be consistent with GeoJSON.
5. Despite the metadata retention being substantial it is not holistic. There are many metadata features not captured in the files as proposed which might be of use to researchers. Has thought been given to how to associate additional free-text / rich metadata with MODF files?

Minor comments

1. Line 28 – ECVs are defined by the Global Climate Observing System and not by the WMO
2. In line 326 what is the section reference in the parentheses to? Or is this a legacy needing removal? Its unclear to me.
3. In Table 2 the final column is completely screwed up with random row allocations that make no logical sense
4. In Table 2 discussion paragraph starting line 371 this should surely be 'lat_sonde' and 'lon_sonde'? It might also be worth nothing whether MODFs can cater for descent data which is increasingly being used and exploited.