Comments on "Observations for high-impact weather and their uses in verification" by Marsigli et al.

General comments

This paper is an important contribution to the literature on forecast verification. It sets the stage for the use of new kinds of "observations" for verification and the application of verification approaches to forecasts of new, often user-relevant, phenomena (e.g., thunderstorm occurrence and impacts). Future work in this area will build on the information provided in this paper.

The paper provides citations and summaries of recent work in this area; however, it tends to be fairly focused on research in the European region, with most of the references also from the European literature. It would be useful to include additional references from other parts of the world where different experiences and knowledge exist, but the European focus is not surprising since many of the co-authors are members of the European scientific community. However, it is worth noting, for example, that U.S. researchers have made significant use of radar mosaics to evaluate convective and precipitation forecasts for the last two decades; just a couple of examples include Gilleland et al. (2009) and Roberts et al. (2013).

Another area of research that might be considered in the paper is work done by Hitchens et al. (2013) to define a "practically perfect" warning region based, for example, on point-based storm reports (e.g., for wind, hail). The method uses statistical methods to convert point observations across space into a field (a "practically perfect" forecast) that can be compared directly to a warning or the output from a model.

A relatively minor – but perhaps relevant – consideration concerns the nuance between the words "evaluation" and "verification". While we do *verification* to derive numbers that represent estimates of performance or skill (for some purpose such as monitoring changes over time), it may be more appropriate to use the word "*evaluation*" to represent many of the kinds of analyses that can be undertaken using the datasets considered in this paper. I wonder if including that terminology in the paper would help express the breadth of effort that is required to understand forecast performance, particularly for phenomena that have direct human consequences.

Minor comments

1. Line 36: I believe you mean "weather-related hazards" rather than "hazards weather-related".
2. Line 39: The end of this line ("weather forecast, one the main") needs editing.
3. Line 61: The last part of this line ("…combining to…") needs editing.
4. Line 76: Can you suggest other phenomena that would benefit from application of these kinds of approaches?
5. Lines 142-145: I want to note that some of the spatial methods (e.g., distance metrics, MODE) do make it relatively easy to identify and/or evaluate false alarms.
6. Line 241: Note that the human impact may still be large even if the population is sparse (e.g., in the US Midwest, localized hailstorms can destroy farm crops and have huge economic impacts).
7. Line 269: It should be made clear that the studies mentioned are a subset of those that have been undertaken world-wide to address this topic and that many other research efforts could be cited.
8. Line 289: Should "P" be "POD"? Also, what about FAR in this example?
9. Line 341: Should this be FAR rather than false alarm rate?

10. Finally, a few typographical and minor grammatical errors, scattered throughout the paper, should be corrected.

References

Gilleland, E., D. Ahijevych, B.G. Brown, B. Casati, and E.E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. *Weather Forecast*., **24** (5), 1416 - 1430, doi: 10.1175/2009WAF2222269.1.

Hitchens, N.M., H.E. Brooks, and M.P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525-534.

Roberts, R.D., A.S. Anderson, E. Nelson, B.G. Brown, J.W. Wilson, M. Pocernich, and T. Saxen, 2012: Impacts of forecaster involvement on convective storm initiation. *Weather and Forecasting*, **27**, 1061-1089.