Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., author comment AC3
https://doi.org/10.5194/nhess-2021-215-AC3, 2021

Reply on RC2

Philip A. Ebert and Peter Milne

Author comment on "Methodological and conceptual challenges in rare and severe event forecast verification" by Philip A. Ebert and Peter Milne, Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-215-AC3, 2021

Thank you very much for this extensive and extremely helpful review. We are very pleased to learn that we make "very valuable" recommendation and so add to the current debate.

It was not our intention to mispresent Heirli et al (2004) and the follow-up study by Purves & Heirli (2006). In our revisions, we will make sure that we represent the aims of these articles more clearly (and correctly!). These two papers directly influenced us in writing this article so we will make sure to give them the credit they deserve.

Now, while the paper is already quite long, we will endeavour to make the paper more comprehensive in the following ways.

- We will expand Table D1 as suggested to include the different labels for the skill scores as they appear in the literature and look through recent articles in NHESS and include references to these in the table.
- We will make sure to cite existing discussions of the Peirce measure, such as Monzanto (2007) (which we weren't aware of) and possibly other work by Bob Winkler that we recently came across.
- As you note there are other research areas in which similar issues arise. For what it is worth, we do already refer to some papers in the machine learning literature when mentioning the accuracy paradox; we will expand our discussion to include Brownlee (2020) — it seems to be the most comprehensive discussion to date. Note that Brownlee recommends the F-measure, common in information retrieval since the 1970s; this is in fact the same as the Dice co-efficient which we mention in our article and which Joliffe advocates in his (2016). In a revision, we will say a little more about why this measure isn't suitable: in short, a) it is somewhat arbitrary (in that it takes the harmonic mean of precision and recall, but why not the arithmetic or geometric mean (both of which have been used in other areas of research)?), b) it ignores 'd', the number of non-occurrences correctly predicted to be non-occurrences, thus making "better than chance" calculations impossible, and c) it fails our second requirement of

"direction of fit", since it doesn't have one. Finally, we will also add related literature from remote sensing (such as the paper entitled "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment disagreement for accuracy assessment" by Robert Gilmore Pontius and Marco Millones (International Journal of Remote Sensing, 2011), which in effect is a rejection of the Heidke Skill score). We hope this will not expand the paper too much and does directly engage with the concerns raised in the last part of your report.

Thank you so much for taking the time to read our work and give us your comments. They will undoubtedly improve our paper and increase its relevance to the wide readership of this journal.