Dear Pascal,

Many thanks for giving us the opportunity to revise our manuscript. Below we note how we meet the comments by the referees and your own comments. We haven't made a note of merely stylistic changes but changes in responses to comments are highlighted in the submitted manuscript.

1. Prose

We reworked our prose. It's difficult to completely change our style, but we tried to be more direct and, hopefully, thereby more informative. As such we cut out repetition of "flagging-points" — see in particular the end of section 3.3 (difficult to highlight deleted passages!) — and trimmed historical detail that didn't move the argument forward. In addition to these deletions, we have made numerous additions (including to the bibliography) — see below for details — so that, all in all, the typescript ends up much the same length as before.

2. Comments from the reviewers

We incorporated all technical corrections and small comments from referees 1, 2 and 3 as we promised we'd do. This includes the comments from reviewer 1 about a confusing sentence concerning the accuracy score, as well as all comments in the pdf by reviewer 2.

3. Reviewer 3

- In light of Reviewer 3 comments to include additional material from machine learning and other related areas, we can note that we added a reference to Brownlee 2020 who also discusses the accuracy paradox. In addition, the most widely used measure in machine learning and in information retrieval, the Dice coefficient, a.k.a. the F measure — also adopted by Brownlee — is now discussed throughout the text (see the end of section 3.1 and 3.2) and added into Table 7. We show that it fails to meet our adequacy constraints and argue that it is therefore unsuitable for RSE forecast verification.
- We added, as requested, references to Manzato 2005, 2007 (see section 3.3).
- We looked for existing literature on skill measures in this journal. Most of the discussion concerns probabilistic forecast-verification which uses other measures that we don't discuss given our focus on binary forecast verification. However, we did find two recent articles in this journal that use the True Skill Statistic (TSS, the Peirce measure by another name) and other measures in the context of landslide forecast verification and landslide categorisation. We made references to these papers in section 4 to highlight that the same kinds of issues arise for landslides.
- We expanded and made more comprehensive our table D1 in the appendix. It now contains all the different skill measures and at least some of the different names for these measures so that our article can be used as a comprehensive reference point for the evaluation of binary RSE forecasting.
- Finally, Purves made a note about not properly appreciating his and his co-authors' two contributions on the topic at hand. We made changes to section 4 to:

- make sure the scope of their claims is better represented, and
- make clearer how much our discussion is indebted to their earlier work.

4. Wider readership

As you can see, we added some work from machine learning, information retrieval and from the landslide literature. We hope this will widen the intended readership. Moreover, as noted above, we expanded our table D1 so to offer a more comprehensive discussion.

Unfortunately, we were not able to expand our discussion to the remote sensing literature as planned. We tried but it added too much additional prose. The remote sensing literature typically uses multi-categorical measures, in particular Cohen's kappa measure, which is in effect the Heidke score. However, we don't mention multi-categorical measures until the end of section 5. Trying to add substantial comment on them into our discussion at the end of the article just didn't work. Trying to add it earlier proved to turn into a page long distraction that we felt wasn't warranted and seemed in context somewhat artificial. As we note in the paper, we plan to do future work on multi-categorical measures and so this discussion about remote sensing literature has to be postponed to that occasion.

Examples from our literature review of skill in this journal that we don't cite are: (https://nhess.copernicus.org/articles/21/1297/2021/, https://nhess.copernicus.org/articles/20/1595/2020/, or other skill measures for probability distributions such as the Jensen–Shannon divergence (JSD) from https://nhess.copernicus.org/articles/20/107/2020/ )