Authors' response to the editor

Thank you for the warm feedback and the suggestions. We have made all of the editor's suggested changes (details below). Please let us know if there's anything else we should do. We are looking forward to seeing our Review and Perspective paper published in GMD.

1) In response to reviewer #2, the authors presented results of a small study of ML publications from the journals BAMS, GMD, and JAMES. I think this is interesting and, because the authors mentioned that they would be willing to do so, I would like to encourage them to present this study in their manuscript. I note that this is optional, but I believe that presenting their statistics helps make clear the significance of their review and perspective paper. Please note that if this is added, the data should be made available and the data/code availability section updated to reflect this.

We have added this to the manuscript. In particular, we added a figure which shows the trends in ML publications at BAMS, GMD, and JAMES over time. We have uploaded the data and a description of the methods as supplementary material.

2) Reviewer #1 asked that several technical terms be better defined. The authors have asked for my judgement on the matter. Of the terms reviewer #1 mentioned, I agree that it would be good to better define or explain "attribution/relevance heatmaps", the distinction between "specific classificatory instances" and "global classification", and "P-score". Regarding that last item, I suggest that, where the authors have written "Indeed, NNU had a far lower P-score in both the baseline and the 4k warmer climate cases", they quantitatively state the reduction in P-scores. Reviewer #1 also mentioned "layerwise relevance propagation". I noticed that this is actually already explained in footnote #3, but the paper needs some reorganization to make this clear earlier, as the term is used in footnote #2 and the main body of the paper before footnote #3 is encountered. Reviewer #1 also mentioned "tree ensembles". I don't think this necessarily needs better explanation, but I might expand this to the more descriptive "ensembles of decision trees", and perhaps also note that a very common example off this is random forests. (I do not think that multi-layer convolutional neural networks needs explanation, as these are very commonly encountered in the literature in the context of geospatial data.)

We now define "attribution/relevance heatmaps" as suggested. We have moved our description of "layerwise relevance propagation" earlier, so readers will see it defined in a footnote on first use. We changed the "random decision trees" as suggested. We clarified

what is meant by “specific” vs “global” classificatory instances. Finally, we fixed the “P-score” description and corrected the relevant discussion (it was supposed to be a physical constraints penalty *P*, meaning that lower scores indicate less violation of physical constraints, given in units W^2/m^4). We’ve added quantitative values of *P* to the manuscript as well.

Minor comment: In line 505 of the author-tracked changes, there is mention of a "certainty deficiency". I believe that the authors mean "certain deficiency"? (Or do they mean a deficiency in certainty? I am not entirely sure.)

Yes, it should be “certain deficiency”. We fixed it – thanks!