Review of “Classification of synoptic circulation patterns with a two-stage clustering algorithm using the modified structural similarity index metric (SSIM)”

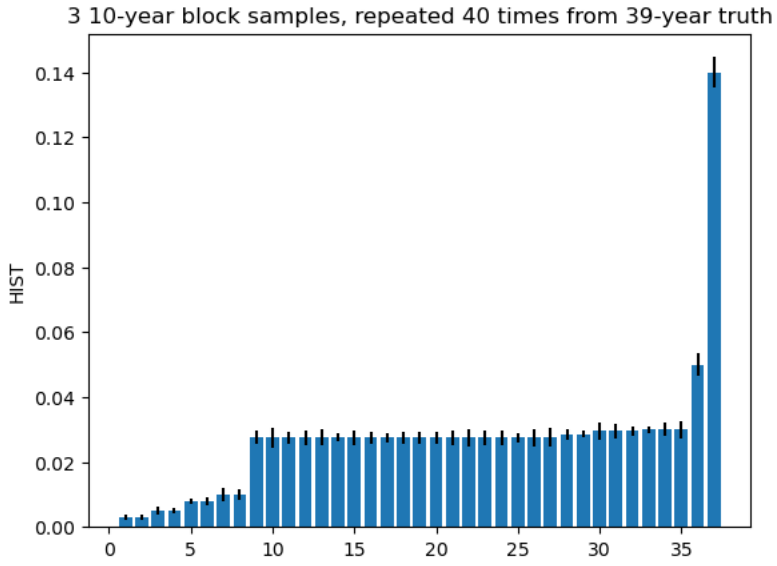
In my previous review I raised some major issues: The claim that a large number of full field regimes would help capture extremes was unproven. The suitability of the JS index for capturing extreme event errors was unclear. The skill metrics used were very granular and it was not obvious to me they could be usefully constrained in observational or simulation data.

All these points have now been addressed, however I am not fully convinced by the treatment of the third point.
Taking the paper’s results as given, the sensitivity testing of the HIST metrics and others to resampling look very good: it is easy to see that the model errors are far outside the 2*std range shown in many cases.

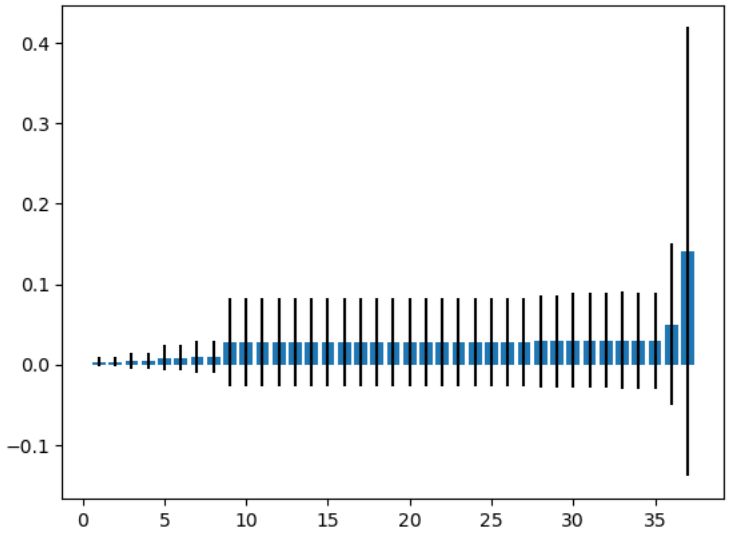
However, I was really surprised to see such small sampling errors. Taking one exemplary case: if we look at class 37 in figure S8, the mean occurrence is around 0.3%, with a standard deviation clearly much lower than 0.1%! I believe these small stds are down to the way the bootstrap is computed (3 ten-year samples), and give a spurious impression of how well the elements are constrained.

To investigate, I coded up a synthetic example, randomly sampling from 37 states with true occurrence frequency:
([0.003,0.003,0.005,0.005,0.008,0.008,0.01,0.01,
 0.0275,0.0275,0.0275,0.0275,0.0275,0.0275,0.0275,0.0275,
 0.0275,0.0275,0.0275,0.0275,0.0275,0.0275,0.0275,0.0275,
 0.0275,0.0275,0.0275,0.0285,0.0285,0.0295,0.0295,0.0295,
    0.03,0.03,0.03,0.05,0.14])

I took a 39*365 day sample as my ‘ERA INTERIM truth’, and then computed sampling errors using the method described in the new manuscript. I get a result very similar (qualitatively) to the authors, plotting error bars as 2*std as they do:

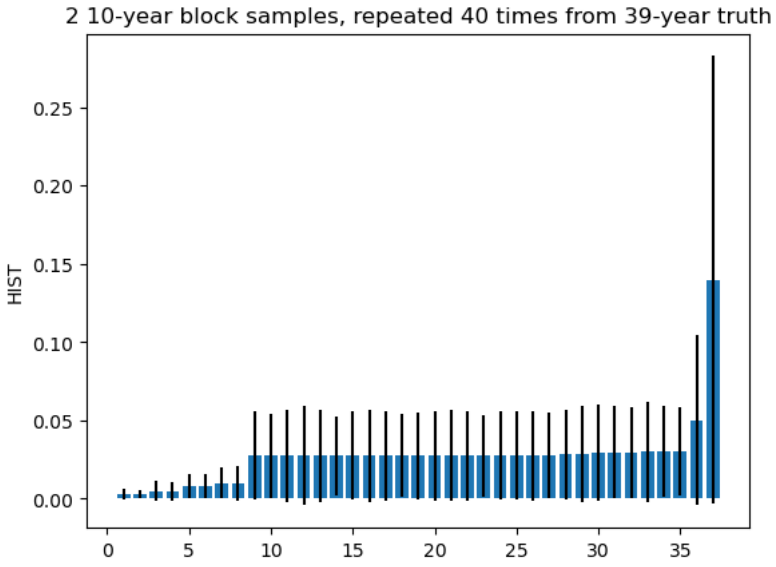


However, if I produce 1000 39-year samples of my state vector, and assess the actual sampling error (i.e. the expected error between two 39 year sampling periods or ERA-Interim and an uninitialised climate model run) I get the following plot:



The bootstrap is clearly massively underestimating the sampling error. Of course the authors can not magically obtain another 1000 years of ERA Interim data, but the point is that sampling 30 years from 39 years does not provide sufficiently independent samples to well represent errors.

Dropping the bootstrap down to 20 years (2 blocks of 10) gives much more reasonable results and I recommend the authors do this and then reanalyse their conclusions. Alternatively they could download ERA5 data and repeat the analysis, as that now covers a 73 year period, and allows for much larger independent bootstraps.



As I consider the current error quantification to be (accidentally) misleading, I am again recommending major revisions. All other aspects of the manuscript are much improved however.