Review of “The influence of the Atlantic Multidecadal Variability on Storm Babet-like events”

The work of Thompson et al. presents an investigation of the link between Atlantic Multidecadal Variability (AMV) and analogues of storm Babet (October 2023). The authors seek to investigate the sensibility of analogues averaging inside time slices as done by Faranda et al. (2022) to conclude on the influence of anthropogenic climate change on similar storms. The goal of the paper is to quantify the impact of multidecadal variability (AMV in particular) in the North Atlantic on the analogues found using multilinear regression on variability indices.

The paper is well-written and the research question is clearly stated. The latter is an important avenue of research for quantifying the sensibility of the analogues method to the long-term variability of the climate system and understand better what kind of conclusions can be drawn from this method as it is now used for attribution purposes (the authors cite the Climameter tool). Moreover, the work of Thompson et al. is a contribution to the literature on the dynamical evolution of the atmosphere under anthropogenic and natural forcings. This makes this work suited for publication in Weather and Climate Dynamics.

However, at this stage I recommend major revisions before accepting the paper for publication. In general, the paper lacks sensitivity tests to convince the reader about the robustness of the results found. But more importantly, the authors seem to claim to have found a causal link between AMV and analogues of storm Babet. I am not convinced that the arguments of the authors are sufficient to defend such a statement. Statements in the conclusion such as L193-194: “For Storm Babet, we show that similar events are more likely during positive AMV phase.” and L196: “The results suggest that if the current trends of amplified warming in North Atlantic sea surface temperatures continue, we should expect to see more events similar to Babet.” are not, or very weakly, supported by the analysis provided here. Moreover, if such statements were to be supported more strongly in a revised version of the manuscript, they should be compared to the scientific literature on this subject. See below for the details of my comments.

Major comments:

1. Sensitivity tests:
    a. The authors decided to use an arbitrary number of analogues in each period. While the number of analogues considered will always be more or less arbitrary, I think the authors should investigate the sensibility of their results to the number of analogues chosen (especially for Fig 1 and 2).
    b. Have you assessed the sensibility of your results to the spatial domain chosen for defining the analogues?
    c. In Fig 1 and Fig 2, how the statistical significance is obtained is not clear. The legend in Fig 1 indicates that it is obtained when the mean anomaly is below the standard deviation. I am not sure to understand what that means and I do not think this is a proper statistical test or procedure.
    d. L155-164: the results of the regression are not reported properly: there is no indication on the statistical significance of the regression, nor on the R^2, nor on the uncertainties and p values for the coefficients for AMV and GMST.

e. Fig 4c: I guess the shadings of the GEV represent the uncertainties. How were they obtained? Why do they begin only for return periods greater than 10 years?
f. Fig 4c: The authors should be more explicit in what they fit here. I guess that the location parameter depends linearly on AMV, but does it also depend on GMST? Please explain more clearly what GEV is fitted and with which fitting method.
g. L173-179: None of the values reported here have a confidence interval associated.
h. Fig 5:
   i. It is not clear to me why these two boxes were chosen. How much does this result depend on the region selected? You could rather show a map with the difference in the mean at each grid point to see how this pattern varies geographically.
   ii. I am not sure I understand the argument behind this figure. Using days for which S>0.7 is not completely equivalent to using days in positive vs negative phases of the AMV. Maybe you could find analogues in the two phases of the AMV and see the difference? This would be similar to what Cadiou et al. (2023) did for ENSO.

2. Impacts of AMV and causality statements:
   a. The authors use multilinear regression to assess the combined effect of AMV and GMST on Sx, a measure of the quality of the yearly best analogue of storm Babet. I will assume here that using Sx is correct (see below for my comment on this point) for estimating the sensibility of analogues. However, I am not sure how we should interpret the results of this regression. The authors seem to say that a significant AMV coefficient is sufficient to conclude to a causal effect. It may be true but the authors do not give enough arguments to support this claim. First, it is clear to me physically that if one finds a link between AMV and Sx, it is probably because AMV influences Sx rather than the contrary. The authors should explain that more clearly because it is currently implicit in their formulation. Second, to correctly estimate the impact of AMV on Sx, one needs to control for **all confounders**, i.e. for variables that have a causal impact on both AMV and Sx, and **only for these confounders**. The authors say in L 215: “We acknowledge that anthropogenic forcings may be influencing AMV, but this will not impact our key findings.” Actually yes it would: any variable that would influence both AMV and Sx (such as anthropogenic forcings) and that is not taken into account in your regression will bias the estimation of the AMV coefficient. On the other hand, here the authors seem to suggest that GMST is such a confounder, which does not seem correct to me. Moreover the authors do not detail and discard other potential confounders (anthropogenic aerosols for example?). I recommend the work of Kretschmer et al. (2021) for clarifying these points.
   b. The correlation between AMV and Sx is rather weak (around 0.5 which means that only 25% of the variance is explained). Moreover, Fig 4b is quite worrying: it does not convince me that there is a linear relationship between these two quantities. If anything, it would suggest a second order relationship

(ie proportional to AMV**2). Finally, the authors project their results for a value of +0.25 for AMV which is far above what is ever observed in the data set. It seems to me quite problematic: how can you be sure that other data points will not have lower values of the similarity for high values of AMV as the negative trend for points on the right of AMV=0 suggests?

c. The key results of this work are based on the analysis of the yearly max Sx. I think this is problematic for several reasons.
    i. I do not see the point of using yearly maxima here. It is not clear how these yearly maxima reflect correctly the analogues found: how many of these maxima correspond to analogues used in Fig 1 and Fig 2? How many analogues do not correspond to yearly maxima? (there could be several analogues in a single year and none in other years).
    ii. The core of the results lies on the non-stationary GEV fitted on the Sx. Although this method is interesting, I am not sure we can interpret the results as straightforwardly as assumed by the authors. What the GEV gives is the probability that the yearly maximum of S is above a given threshold. It does not say how many analogues there would be per year and whether they will be good: it only gives a probability to the best one being very close. The authors say for example in L174: "events similar to Storm Babet are more likely – the chance of Sx>0.7 is 7.5 times more likely during AMV positive than negative", I am not sure this assertion is supported by the fit on Fig 4c because I do not think the fit of the GEV on the Sx can be straightforwardly linked to neither the quantity nor the quality of the analogues in the positive and negative phases of AMV.
    iii. It seems to me that for what the authors want to achieve, the GPD approach for the extremes of S is more suited because it takes into account the quantity of analogues in addition to the shape of the tail. I want to highlight the fact that using a GPD (or GEV) approach on the distance of the analogues of a point (ie linking extreme value theory and recurrence in dynamical systems) is a mathematical domain that has been well developed in the recent years (see Lucarini et al. 2016 for an exhaustive review) and that it is the mathematical foundation for the local dimension measure provided in Faranda et al. (2022).

Minor comments:

1. L27: typo "stormtims"
2. L95: maybe the number of the section is missing
3. L104: this equation should have a random term epsilon
4. L179: using the Mann-Whitney test to find differences in the distributions is okay but here you report differences in the averages of these distributions. A Welch's t-test seems more appropriate to me.
5. Figure A2: the color of the 95th percentile is indicated to be orange but is green in the figure. Also, why are there no units on the x and y axis of panels b and c ?

References:

Kretschmer, M., Adams, S. V., Arribas, A., Prudden, R., Robinson, N., Saggioro, E., & Shepherd, T. G. (2021). Quantifying causal pathways of teleconnections. *Bulletin of the American Meteorological Society*, 1-34.

Lucarini, V., Faranda, D., de Freitas, J. M. M., Holland, M., Kuna, T., Nicol, M., ... & Vaienti, S. (2016). *Extremes and recurrence in dynamical systems*. John Wiley & Sons.

Cadiou, C., Noyelle, R., Malhomme, N., & Faranda, D. (2023). Challenges in attributing the 2022 australian rain bomb to climate change. *Asia-Pacific Journal of Atmospheric Sciences*, *59*(1), 83-94.