Review – *Attribution of the Australian bushfire risk to anthropogenic climate change* by van Oldenborgh et al.

This paper describes an attribution analysis of a number of factors that are known to either contribute to or reflect wildfire risk using observational data, observationally constrained data products (reanalyses), and a collection of CMIP5 model simulations. It also considers the impacts of internal climate variability as reflected in the large-scale modes of variability that influence the Australian climate, and it includes a discussion of vulnerability and exposure factors associated with the impacts of the summer 2019/2020 wildfires.

I found the paper frustrating to read and evaluate. One clear impression is that the authors were in a terrible hurry, producing text that often appears not to have been carefully proofread, not thinking carefully about how to describe their methods in a clear way, not always justifying methodological choices, not justifying choices of data products or evaluating those products with a sufficiently critical eye, and attempting to be overly comprehensive. Reading the paper is a bit like being forced to "drink from a firehose" – there are so many details and so many small aspects that can be criticized, that is difficult to know exactly how and what to criticize in a review. The fact that all code is being made available doesn't really reassure me very much. Readers who want to understand what was done, sufficiently so that the work can be replicated, shouldn't be placed in a position of having to read code but rather, should be provided with explanations in the paper that are clear enough so that they can develop and implement their own code.

Some specific comments:

1-14: The abstract does not mention the long section on vulnerability and exposure factors, and there is no reference to vulnerability and exposure in the title. Does that section really belong in the paper?

16: The very first sentence of the paper starts by being sloppy in the way in which Australian station data are characterized. The word "homogeneous" has a very clear and well understood meaning in the context of observational data products (i.e., meaning that observations have been carefully evaluated and adjusted to ensure that they are free of artefacts resulting from changes in instrumentation, instrument siting, instrument housing, observing and reporting practices, etc., etc.), and surely the claim here is not that Australian station data is homogeneous in that sense. Clearly, avoiding the obvious inhomogeneity due to the lack of proper instrument shielding early in Australian instrumental record is necessary, but we shouldn't just accept that all of the subsequent record is homogeneous.

26-27: What is the source of this estimate? Is it possible to have any confidence in that number or the range that is given?

27-29: Again, what are the sources?

Figure 1: Is there a URL and a date for where this image was obtained?

93: I imagine daily maximum temperatures are meant. There are many instances in the paper where a second reading of the words, just to see if the connect logically, would have helped enormously. There are also a large number of run-on sentences in the paper that are difficult for readers to parse and understand.

102: This subsection is entitled “Event definition”, but it doesn’t talk specifically about event definition at all. I think what is needed is a clear statement that the event of interest will be defined using the FWI. This section gives some justification for doing that by considering the relationship between FWI and area burned, but event definitions per se are not discussed in this subsection.

Figure 2 caption: Please tell me what is meant by a “one-sided confidence interval about zero”. I assume you mean the interval from -1 to the expected 95^{th} quantile for the correlation coefficient under the null hypothesis that the correlation is zero. If this is correct, then it would be better to call this the 5% significance critical value for a one-sided test of the null hypothesis that the correlation is zero against the alternative hypothesis that the correlation is positive.

129: Often, acronyms like ASF20C appear before they are defined.

137-152: Some careful justification for the distributional choices would seem to be in order. These distributions emerge in statistical extreme value theory as limiting distributions under idealized conditions, where the limit is taken either as block length increases without bound in the case of the GEV, or as the exceedance threshold increases without bound in the case of the GPD. Given the way the data are processed, we are likely a long way from being able to be satisfied that the actual distributions are well approximated by these limiting distributions. Indeed, it seems likely that the relative quality of the fit will diminish as you go deeper into the tail, even if quantile plots look to be ok. In particular, one should be worried about extrapolating beyond the available data. Some aspects of this are discussed later in the paper, but those limitations don’t really seem to prevent the authors from referring to values that appear to correspond to very long return periods in some instances. In the case of precipitation deficit, any of a number of possible candidate distributions could presumably be considered if using as much as 30% of sample values. These would have different deep tail characteristics, affecting calculations of probability ratios, but might not be discernably better or worse than the GPD based on standard diagnostics of the fit. So how does one proceed in a careful way take this source of structural uncertainty into account? It might be as important as the structural uncertainty represented by the spread between models.

160: Why 4-years and not some other degree of smoothing? Exactly how is the smoothing done, and how is time referenced to the smoothed values? For example, if using a 4-year running mean, which year is the value associated with in covariate dependent functions?

165-172: Choices for how the GEV and GPD distributions are parameterized should be justified and carefully argued, not just stated. For precipitation, exponential scaling might make sense at the upper end of the precipitation distribution, but why would I consider that to be reasonable at the lower end of the distribution, and why, in that case, should the scale parameter be linked to GMST? Building in something that scales like Clausius-Clapeyron might not be the best idea for the dry end of the precipitation distribution.

187-189: Is it obvious that this is the best way to proceed? If the analysis was literally performed as described here, the effective block size for the models would be 5- or 10-times the block size used for the observations. That means that for the models, the block maxima used to fit GEV distributions would sample a much deeper part of the tail than is possible with the observations since the distributions for the model output would have been fitted to what are effectively 5-year or 10-year blocks rather than 1-year blocks as for the observations. How then, can I make sense of differences in parameter estimates between fits to observed and simulated parameter estimates.

191-192: I think this is all that is said about bias correction in the paper except for another brief mention at line 442, but surely this is important and should be discussed (and defended) in some detail. Exactly what was done, and how does this avoid overusing the observational data?

200: Exactly what do you mean by the χ^2/dof statistic (what is calculated, and what is the basis for the interpretation given to this statistic)?

240: For each observational product, the paper should draw attention to the key limitations that would affect the analysis in this paper. For example, although Stevenson screens begin to be used in 1910, there could be many other reasons to be concerned about the homogeneity of temperature observations, such as variations in station coverage over time (e.g., spatial sampling in 1910 would undoubtedly have been different than in the 1970s). Also, the paper should make a clear distinction between observational products on the one hand, and observationally constrained products (re-analyses) on the other. The latter are clearly non-homogeneous, with inhomogeneity due to changes in data sources, quality and quantity over time being of particular concern in the southern hemisphere where the observational constraint is much weaker. Ensemble reanalysis products, such as the 20th century reanalysis may be able to provide information about the strength of the observational constraint and how it varies in space and time (if the spread between ensemble members is large, the constraint is obviously weak or non-existent; if the spread is small, one has further work to do to determine if it is small because the analysis is being effectively constrained by the observations or whether this is coming about for another reason). Further, it should be noted that surface variables are often not very well constrained in reanalyses. The classification of variables by strength of observational constraint that is given in Appendix A of the Kalnay et al paper describing the original NCEP 40-year reanalysis (BAMS, 1996, *https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2*) still largely holds and should be considered.

Figure 3: Use the same vertical scale on both panels (or better yet, plot the two timeseries on the same graph).

251-255: A number of reanalysis products are mentioned here, but the paper also uses others (e.g., ERA-5).

240-269: An overview of the strategy for using the different observational and reanalysis products would be useful. This would demonstrate that there is some overarching reasoning that knits the selection of products together and that has informed the choice of products. I have to say that the choices are really confusing, both for reanalyses and for the observational products. For example, GMST is apparently from GISTEMP (mentioned at line 123, but not in this observational data section), but the gridded global surface temperature dataset that is used is Berkeley Earth (line 242), and other well studied and documented global gridded temperature data products such as HadCRUT4 are not mentioned at all. Why these particular choices? For the gridded products, the infilling strategy and error models, which vary between choices, are presumably important considerations, particularly in the southern hemisphere and especially when considering a relatively small land area in the southern hemisphere that is sandwiched between ocean to the east and a very dry, sparsely observed continent to the west.

270: I find it very surprising that the entire observational discussion for TX7x, including results from AWAP and mention of one of the reanalyses, is limited to only 6 lines of text. Statistical model fitting results are shown in Figure 4, but are really not discussed in any meaningful way – and Figure 4 itself is not explained in a way that most readers would be able to understand. Specifically, cumulative frequency distributions for 1900 and 2019 are shown, but there is no explanation in the text or in the figure caption explaining how the points that are shown are derived from the observations. Evidently observations are adjusted to particular years using smoothed GMST values for those years to make adjustments via the fitted distribution. Shouldn't one be concerned that this could induce some circularity, particularly if one of the intents of the figure is to illustrate the fit of the statistical model to the observations? Results from one reanalysis are mentioned, but silence concerning other reanalyses begs a question about whether they did not "tell a similar story" – do they tell a similar story?

281: See my comment concerning lines 187-189. What explains the apparently much narrower uncertainty bounds on the climate model-based parameter estimates as compared to the model-based estimates? Is the explanation that the model-based analysis actually uses annual blocks rather than blocks constructed by pooling data for a particular simulated "year" across ensemble members (which is literally what lines 187-189 appear to say)? In this case, samples of annual maxima are 5- or 10-times as large as from observations, which, all else being equal, should result in confidence intervals that are about $5^{-0.5}$ or $10^{-0.5}$ as wide as for observations (i.e., ~45% or ~32% as wide, respectively). But this interpretation also doesn't seem quite right because the model confidence intervals seem narrower than these expectations.

I have many largely similar comments about sections 4 and 5 that I won't repeat here. Hopefully the message that the paper needs to document the work and justify choices and interpretations much more carefully has come across.

Regarding Section 6 – a very strong conclusion is drawn on lines 565-566, but it is not obvious to me that the strong quantitative evidence and supporting modelling experiments that would be required for such statement has really been presented. Quantitative evidence seems to be restricted largely to estimates of correlation coefficients which, if considered as simple regression diagnostics (i.e, focusing on r^2 rather than r), would correspond to explained variance amounts of the order of 5-15%.