Response from Allen

I would have expected to have to produce a reply to Ross McKitrick's paper before it appeared in the journal – this is the normal procedure for comments on articles like this one. Professor McKitrick only told us about it days before it was published, and the journal didn't contact us at all, which is very odd.

My general thought is that fully addressing the issues raised by this paper might have made some difference to conclusions regarding human influence on climate when the signal was still quite weak 20 years ago, although he doesn't show whether resolving these issues would have led to human influence been detected earlier or later than it was. But this point is now rather moot, since the signal is now so strong that you get essentially the same answer for the estimated amount of warming due to human influence whether you use a relatively elaborate method – as in AT99 – to maximise signal-to-noise or a simple ordinary least-squares regression on global average surface temperature.[*] The original framework of AT99 was superseded by the Total Least Squares approach of Allen and Stott (2003), and that in turn has been largely superseded by the regularised regression,[†] or likelihood-maximising approaches,[‡] developed entirely independently. To be a little light-hearted, it feels a bit like someone suggesting we should all stop driving because a new issue has been identified with the Model-T Ford.

Most of the points he raises are already quite familiar: indeed, several are discussed in AT99 itself. His main issue is with Gauss-Markov condition (3), which is that the noise *u* is independent of the signals being searched for, P_X. It is very important that a statistically independent realisation of noise is used for hypothesis-testing than is used for the estimation of the pre-whitening operator P: failure to do this introduces a bias, as already recognised by Hegerl et al. (1996), and discussed in AT99.[§] Stripped of the jargon, we need to use one source of information to define what we are looking for, and then another completely independent one to quantify how surprised we should be at having found it. McKitrick doesn't (as far as I can tell) discuss this point at all, so he may have simply missed that this is what is done, which is unfortunate.

In response, McKitrick would probably argue that even if an entirely independent model run is used for hypothesis-testing, any climate model used 'embeds the assumption that greenhouse gases have a significant effect on the climate', as he puts it (rather disingenuously: the fact that greenhouse gases have a significant effect on climate is not an embedded assumption in the kinds of climate models he is talking about, but emerges from the way these models behave). There is a real issue here, which has long been acknowledged in the detection and attribution community,[¶] but it means that conclusions based on the AT99 approach would actually be conservative, overstating uncertainties rather than the opposite.

Everyone (including McKitrick) accepts that increasing CO_2 has a predictable impact on the global energy budget; between 3.5 and 4 Watts per square metre for a CO_2 doubling. So the uncertainty is in how much the climate system needs to warm up to compensate. According to the fluctuation-dissipation theorem, a model that displays a low sensitivity to rising greenhouse gases because it dissipates energy very efficiently per degree of warming would also display lower internally-generated temperature variability, for precisely the same reason. Hence, in using models with climate sensitivities in the standard 2–4°C range to test the null-hypothesis of a zero climate sensitivity, we are likely to be over-estimating the noise variance and hence the uncertainties. By how much is a moot point, since efforts to link variability with climate sensitivity using the fluctuation-dissipation theorem have been controversial, to put it mildly.[**] This is not a new issue: indeed,

* See globalwarmingindex.org and the spreadsheet enclosed therein, for example.

† Ribes and Terray, 2012; Hannart, 2016.

‡ Ribes et al, 2017.

§ See p. 423, column 1.

¶ See the discussion in Ribes et al., 2017.

** See e.g. Cox et al, 2018, and various matters arising.

it was part of the motivation for using perturbed-physics ensembles to provide confidence intervals on the climate response†† rather than relying on scaling a single model's response.

The issues he raises with the residual consistency test all seem to be based on a massive overstatement of the weight given to a single statistic: both in AT99 and ever since, we never suggested that the RCT for a single truncation level was the only number people needed to look at to be confident their climate model-simulated variability was realistic: we explicitly recommended looking at its behaviour across a range of truncations, and complementing it with other diagnostics, such as power spectra.

So, in summary, if we were still in 1999 and dealing with a signal that had only recently emerged from the noise of internal variability, then the issues McKitrick raises might be important, although the main one suggests that, if anything, the approach of AT99 was overly conservative, although by a difficult-to-quantify amount. But we aren't. The signal of human influence on climate is now so strong the fine details of the approach you use to estimate it barely matter. A high-school-level regression analysis of global mean surface temperature,‡‡ gives much the same answer as much more elaborate methods, including both those using climate models and econometric methods that explicitly avoid relying on climate models.§§ McKitrick raises some helpful theoretical points, but makes no attempt to quantify how much difference addressing the issues he raises might make to conclusions, even given the information available in 1999, much less today.

†† See e.g. Murphy et al, 2004; Stainforth et al, 2005; Rowlands et al, 2012.

‡‡ As in Haustein et al, 2017.

§§ e.g. Chang et al, 2020.