**Review of Lynch et al. "Exploring global surface temperature pattern scaling methodologies and assumptions from a CMIP5" submitted to Geoscientific Model Development Discussions**

The manuscript sets out to evaluate some specific elements of the pattern scaling (PS) approach to generating climate change projections. The PS approach attempts to approximately emulate GCMs by separating climate changes into a global-mean temperature (GMT) change and a pattern describing local changes across the globe that scales linearly with GMT. The performance of PS depends to a large degree on the assumption that the same pattern is applicable across a range of scenarios (different rates and amounts of warming and different individual forcing factors). The manuscript ostensibly addresses a different, though closely related issue, namely how to diagnose the pattern for a particular GCM and how sensitive are the patterns to the selected method of diagnosis.

While this is a reasonable aim, the manuscript is not suitable for publication at present and indeed requires some redesign and refocusing from the beginning – thus major and fundamental reworking is needed.

Overall, the major issues are:

1. Given that the point of PS is to explore the spread of projections and to do so at local/regional scales, there is far too much on the ensemble mean and too much on the global mean temperature changes. This is mostly irrelevant to PS. The global mean temperature is an input to PS, rather than an output, so it is very odd to spend so much time analysing it and few conclusions seem to be drawn from it.

2. The work lacks a clear and logical overall framework. It needs to consider what factors will influence the patterns diagnosed, how will each diagnosis method be affected by these factors, how can the pattern differences be explained by these factors? Factors to consider are (i) internal variability versus the signal of forced climate change, (ii) nonlinear dependence on GMT and (iii) scenario dependence. Then consider how method choices affect these: do you use initial-condition ensemble means for each model or just take a single run (this isn't stated), do you apply a time-filter (e.g. running mean or non-overlapping means) before applying the regression approach and how long is the running mean? Then consider how these factors may explain the different patterns found. For example, in the final paragraph of the conclusions it states "the GMT temperature sensitivity is stronger when using a lower forcing scenario because... changes in GMT have a stronger effect on local temperature". This cannot be true across the board because the global mean of the local temperature changes has to equal the GMT change regardless of level of forcing, by definition. Therefore this explanation fails, and it should be reconsidered in terms of what regions exhibit stronger (apparent) local sensitivity between scenarios and why is this more apparent for regression versus delta methods?

3. There are significant flaws throughout the manuscript, possibility suggesting some fundamental misunderstanding of pattern scaling or at least they could lead to misunderstanding by the readers. Some of these flaws are listed below.

(a) The units of Figs. 4, 6, 10 and 11 are incorrect (and units for SI Fig. 2 are not given), given as °C while it is evident from equation (1) for the delta pattern and equation (3) for the regression pattern that the patterns ($DP_{MS}$ or $BETA_{MS}$) are dimensionless (or equivalently are °C/°C, expressing local temperature change per degree of GMT change). This error casts doubt on all the PS patterns shown and analysed here.

(b) The basis of the manuscript is the assumptions that underlie the delta and regression methods used in the literature to diagnose the PS patterns (note that Osborn et al., 2015, is listed incorrectly here as using the delta method but it uses the regression method, explained in great detail). Where the stated assumptions come from is not properly explained and the stated assumptions are in fact not all made by the pattern diagnosis methods. Under "**Assumptions**" it is stated that the delta method assumes that anthropogenic forcings do not modify internal climate variability. This may be an assumption of how PS is subsequently applied *in some cases* to produce a future projection (something that is not considered nor explained in the current manuscript), for example simply adding a PS change to an observed timeseries, but PS does not have to be applied this way (see Osborn et al., 2015, for a PS application using a GCM prediction of enhanced internal climate variability under anthropogenic forcing) and it is certainly not an assumption of the way in which the pattern is diagnosed from the GCM data in the first place. Later it is stated that the delta method to construct a pattern assumes that the trend within the (e.g.) 30-year period is the same regardless of epoch (or forcing scenario). No, it doesn't. See your equation (1): the delta method is simply the ratio of two mean differences, and means over epochs can be computed regardless of whether there is a trend during the epoch. The conclusions make a further related claim, with no support: "the delta method assumes that there is no observed trend in the historical simulation". Where does such an idea come from? If true, it would invalidate the delta method in almost all cases, since nearly all GCMs simulate a warming trend over the historical period. The "**Assumptions**" section then correctly states that the regression method assumes that local changes scale proportionally to GMT and that this relationship is stationary over time. This is correct, but the implication is that this is particular to the regression method – it is not, it applies equally to the delta method and indeed it is really an assumption of the PS approach itself and not additionally an assumption of the method used to diagnose the PS pattern. This is followed up by two unjustified statements: "Transient forcing is likely to scale the local temperature sensitivity to the trend in global mean temperature" (doesn't make sense and is not generally true) and "For temperature-related variables the assumption of stationarity is valid" (not necessarily so, e.g. local temperature change over areas of sea-ice retreat or snow cover retreat). The manuscript then rather hopefully claims that "the differences between the two methods are clear" despite the confusion sown by the errors detailed above – and indeed it is rather a forlorn hope since there aren't any major differences in the assumptions that underlie these pattern diagnosis methods. The underlying assumptions are those of the PS approach itself, which apply equally to both pattern diagnosis methods. The differences will arise because deviations from these assumptions will affect different time periods, simulations and calculations differently and thus the diagnosed patterns will depend on these choices. If these are systematic effects that can usefully guide best practise, that would be interesting – but the manuscript says nothing about this. Further differences arise because some methods have a stronger signal-to-noise ratio than others – but again the manuscript says nothing about this.

(c) The work needs to distinguish assumptions and errors in pattern diagnosis from assumptions and errors in applying the PS approach to generating future projections. This requires consideration of how to assess PS performance, which is lacking. Yes, there are differences between patterns, but which give the closest emulation of the GCM simulation? Significance tests are claimed to show where the linear fit is poor, but the test does not discriminate between poor linear fit and a good linear fit for a weak relationship (e.g. a region where there is little warming in the GCM simulation may have an insignificant relationship with GMT). The first EOF in global annual temperature *may* give the warming trend, but this does *not* require GMT to have a linear trend, and anyway it is not necessarily so (this would depend on the strength of the forced signal relative to the internal variability among other

things) so this statement is incorrect and it is unclear what the purpose of the EOF analysis is. The pattern differences and scenario differences sections (which are rather brief, but presumably are the main purposes of the paper) make a number of comparisons and find a number of differences, but it is unclear what is being shown and what the interpretation is. For example, some high latitude differences are put down to Arctic amplification – but this is an inadequate explanation, since this is in the GCMs and can be captured by PS if it is linearly related to GMT (just with a coefficient > 1). The comparison with the GCM output is unclear and possibly not independent: comparing the PS trend patterns to the GCM trend patterns for the same simulations from which the PS pattern was diagnosed will give an overly optimistic view of the performance of the PS regression method. If the purpose is to establish whether PS based on delta patterns performs better or worse than PS based on regression patterns at emulating GCM projections, then more thought needs to be given to how performance is measured.

(d) There are some further issues (e.g. Fig. 1a is inconsistent with Fig 9.8 of IPCC AR5 WGI, which gives 1961-1990 observed GMT, from Jones et al., 1999, around 14 °C and GCMs scattered around it – whereas Fig. 1a here has annual mean GMT around 10-11 °C for 1961-1990) (e.g. confused use of "variability" – unclear if it means temporal climate variability or ensemble variability/inter-model spread) but of diminishing importance.