Response to reviewers
gmd-2016-170
Lynch et al.

We thank all referees for their thoughtful comments, which have substantially strengthened this manuscript. Per the _GMD_ instructions, we respond below to all comments and questions.

*Original reviewer comments in italics.* Responses in blue. Revisions in maroon.

---

Reviewer #1:

*1. I start from a fundamental disconnect in the interpretation of the regression method. The authors claim that after deriving the beta coefficients they go on to compute the 21st century trend by multiplying it by the number of time steps (page 5, line 14). I don't think this is correct. The beta coefficient, by construction, fits the local trend to the global average temperature time series, so the 21st century trend is only recovered by multiplying the beta coefficients by the global temperature time series, and then fitting a linear trend to the result. So I start by wondering if the entire comparison here is flawed from the beginning by this misinterpretation of what beta is. After all, beta should allow the user of the pattern scaling approach to simulate any new scenario, given its global average temperature time series derived through a simple climate model, while if one followed the authors' recipe any scenario projection would look the same as long as it takes place over the same number of years*

Response:
We agree with the above paragraph. We also agree that the line cited above is confusing because it is incorrect. The regression coefficient (pattern) must be multiplied by the scalar, then divided by the time steps (from the scalar) to obtain a trend per year. But for this case, we are comparing patterns and not the linear time series of scaled patterns. To add clarity to the methods, this line (discussion of construction of a time series) has been taken out as a scaled time series is not used in this analysis.

Revision:
Delete lines 14-15 on page 5.

---

2. *The statistical analysis of differences using a t-test is also open to discussion. The authors use some jargon about a "2-tailed Student's t-test probability using the incomplete beta function ratio" (page 5, line 16/17). I have no idea what that means. The following sentence is also rather opaque (like the majority in this paper, I'm afraid). I found myself sprinkling question marks to the side of almost every paragraph.*

Response:
We agree that this was unclear. The 2-tailed Student's t-test statistic was used to determine where the patterns produced from each method are significantly different. We use this statistic because the ensemble for each method is small, and the ensemble pattern distribution is assumed to be normal The t-distribution can be calculated in different ways, and we used a function in the NCAR command language (NCL) called "ttest", which uses the "incomplete beta function". Some of the details we provided on this function added more confusion than clarity, and we have removed them from the revised manuscript..

Additionally, in the Supplementary section, we added a sentence on the use of NCL for plotting and analysis.

Revised text:
page 5, line 16/17:
"To examine the assumption that the ensemble probability distribution between patterns generated by each method are the same, we calculate the Student's t-distribution probability. This was done because the ensemble consists of only twelve models, a small portion of all available models, and we assume the ensemble variance for each pattern is the same. The resulting probability indicates where there is a significant difference between patterns generated by each method. "

---

3. *I am doubtful that the comparison of regression and delta method (even leaving aside the problem of misinterpretation of the beta coefficients) is justified the way it is carried out through the paper. Why is the comparison always between the regression method that uses the 21st century and the delta method that uses the 19th century baseline? For cleanness, wouldn't it make sense to use the delta method that uses the 20th Century baseline, which by construction addresses the change over the 21st century rather than the change over the 21st AND the 20th century?*

Response:
We initially used the late 19th Century reference epoch for the delta/regression method comparison because that is the closest to the 21st trend, as shown in Supplementary Figure 4. However, after some consideration we think it is important to show difference between different reference epoch patterns to show that epoch choice changes the sensitivity of the pattern, but not the spatial structure. For consistency we have edited the plots to include both reference periods, the late 19th Century and the late 20th Century.

Revision:
Added plot to Figure 8.
Added plot to Figure 10
Added plot to Figure 11

---

4. *What is the point of including the discussion of reanalysis trends? It confuses the reader in thinking that pattern scaling has anything to do with observations (or reanalysis, for those that do not like to think of the latter as observations).*

Response:
We agree with the reviewer that the discussion of reanalysis trends does not fundamentally add to the pattern scaling work. There were 2 reasons for discussion of the reanalysis:
1. It is the basis for the model selection. Out of the ~50 CMIP5 models, we wanted to show that the models we selected strongly agreed with a reanalysis.
2. We wanted to show that for the 19th/20th century there were no clear/strong trends in the models or reanalysis, which would lead one to think that the choice of epoch would not be important. Also, it has been argued that model bias can be ignored because pattern scaling uses the climate change pattern not bias in absolute temperature (Osborn et al, 2015).

We would prefer to keep the discussion of reanalysis, but we agree that it is supplementary and have moved both the discussion of NCEP/NCAR and Figure 1 to the Supplementary material. Discussion of model culling has also been moved to the Supplementary material.

Specific revisions:
Page 4, lines 2-10, are moved to Supplementary material
Page 4, lines 20-27, are moved to Supplementary material
Page 6, lines 3-11, are removed.
References in removed paragraphs are moved to Supplementary material
Figure 1 is moved to Supplementary material and becomes Figure S1.

---

5. *Starting from the abstract: "Temperature patterns generated by the linear regression method show a better fit to global mean temperature change than the delta method" How is the fit to global temperature change of the delta method defined (and therefore comparable)? "global mean temperature sensitivity is higher" (in lower forcing scenarios). What does that mean?*

Response:
The original text was not worded clearly. Page 1, lines 6-7 are in reference to Figure 8, which shows the change in temperature when global mean temperature change is 1 degree as compared to patterns generated by each method. Page 1, lines 8-9, is in reference to Figure 10, which is the difference in patterns generated by each method for each forcing scenario. By sensitivity, we meant the strength of the relationship between GMT change and local temperature, which Tebaldi and Arblaster, 2014, "found that patterns from different scenarios were highly spatially correlated, and that choice of scenario did not explain a significant proportion of variability in patterns when using the delta pattern scaling method." Page 3, line 15-16.
Additionally, t-distribution plots were added to Figure 10 to show locations of significant differences between scenarios for the two delta and the regression patterns.

Revised text:
Page 1, lines 6-10: Differences in patterns between methods and epochs are largest in high latitudes (60-90 degrees N/S). Projected global mean temperature change is more similar to temperature patterns generated by the linear regression method than the patterns generated by the delta method. However, differences are larger across scenarios in the regression method as compared to the delta method, indicating local temperature sensitivity to global mean temperature change is not the same across scenarios. These patterns will be used to examine feedbacks and uncertainty in the climate system.

---

6. *Introduction Why start with the RCP discussion and then talk about computational costs of running scenarios (when all RCPs were in fact run in CMIP5?)*

Response:
According to Taylor et al, 2012 (DOI: http://dx.doi.org/10.1175/BAMS-D-11-00094.1), only 2 (out of 4) future forcing scenarios were considered "tier 1" which allowed modeling centers to prioritize which future scenarios they could run.
Additionally, owing to the complexity of fully coupled models, sensitivity studies or uncertainty quantification are limited. Emulators of coupled climate models are effective in exploring uncertainty by vastly reducing computational cost. Emulators, like scaled

patterns allow integrated assessment models or impact assessment models to include a climate component and incorporate feedbacks between climate and socio-economic sectors. However, the discussion of RCPs in the introduction does not establish a strong foundation for our study, and as such it has been removed and the introduction section has been strongly reworded and reorganized.

---

7. *"scaled scenarios are used for reducing uncertainty". How is that? In fact, the use of scaled scenarios, one could argue, increases our uncertainty by allowing us to use a larger range of futures than would available otherwise.*

Response:
We were not clear in which type of uncertainty was being addressed on page 1, line 19-20. Uncertainty in future climate change comes from three main sources: emission pathways, model response, and internal variability. Pattern-scaling methods allow a wider range of possible future emission pathways and climate sensitivities, thus reducing uncertainty. They could also be used to examine model response uncertainty in regional projections, as done by Dessai et al, 2005 (doi:10.1029/2005JD005919). The introduction has been edited to make this clearer. This is also discussed in the conclusion.

---

8. *The description of what pattern scaling is and how it relates to SCMs is not clear at al, it never mentions the idea of estimating patterns from available GCM experiments and running a SCM to derive a global average temperature time series, whish is the whole point of pattern scaling. A reader less than familiar with pattern scaling would be completely in the dark as to what pattern scaling is and how it works in providing alternative scenarios for analysis.*

Response:
We were not clear the creation of a scaled pattern from previous studies and current software. An additional sentence on Page 2, line 2, will be added. However, we do not believe going into additional detail on construction of a time series generated by a SCM from a scaled pattern would be beneficial. The time series generated from a pattern would be linear unless the addition of temporal variability is incorporated. If using pattern scaling as a method to downscale, temporal variability could be incorporated through the use of analogs or perhaps bias correction. A 2 or 1 way coupling of the scaled patterns from a SCM to an IAM would allow for a time series to be constructed through feedbacks, which is the subject of future work, but will not be added in the conclusion/discussion at this time.

Revision:
Under the assumption a climate variable from a GCM scales linearly with global mean temperature change, patterns are derived from a GCM. Those patterns can then be scaled in magnitude by the global temperature obtained from a simple climate model (SCM) to span a wide range of future scenarios that have not been simulated by full GCMs but emulate the more complex behavior of GCMs (Moss et al, 2010).

---

9. *Assumptions I have big troubles recognizing these assumptions are driving pattern scaling. Pattern scaling by definition does not address internal variability at all. Also, "in practice, estimation errors introduced through this assumption are small". Who says so? This needs a reference.*

Response:
We are addressing the assumptions involved in the methodology to create patterns used in pattern scaling, not the theoretical application of patterns, as this paper is primarily about methodology. We appreciate the reviewer pointing out that this was not entirely clear.

On page 2, line 20, we have qualified that statement by saying at the global scale, errors due to internal variability are small, but at regional and local scales, these errors can be very large (Lopez et al, 2013).

Revision:
This premise is known to be false, but in practice, estimation errors introduced through this assumption at the global scale are small but can be large enough at the regional scale to mislead adaptation decisions (Lopez et al 2013).

---

10. *"In the linear regression pattern scaling method, the underlying assumption is that local change scales proportionally with global mean temperature change and that the relationship is stationary over time". This is actually the assumption of pattern scaling, period, not of the linear regression method uniquely.*

Response:
Page 2, line 33: We agree and have clarified this point.

Revision:
In the linear regression pattern scaling method, as with the delta method, the underlying assumption is that local change scales proportionally with global mean temperature change and that the relationship is stationary over time.

---

11. *For temperature related variables the assumption of stationarity is valid,…" except just a few lines later you say "when pattern scaling patterns of temperature extremes, the magnitude of the error in the pattern estimates was substantially large". Aren't temperature extremes temperature-related variables? And in any event, the statement "the assumption of stationarity is valid" is open for debate. It may be an ok assumption according to some error metric, maybe.*

Response:
We agree that these statements are confusing. We have revised the text to distinguish between temperature-related variables and temperature extremes.

Revision:
For temperature-related variables the assumption of stationarity is generally valid, but the magnitudes of estimation errors vary between scenarios for non-temperature variables (Frieler et al. 2012) and temperature extremes on the upper tail of the temperature distribution (Lustenberger et al, 2014. Lopez et al (2013) found that when pattern scaling patterns of temperature extremes over Southern Europe, the magnitude of the error in the pattern estimates was substantially large.

---

12. *"We use a simplified approach for each method to assess the difference in pattern strength and skill" what is pattern "strength" here? And what is simplified (or simplistic, the way it is worded later in the paper, which I find a little negative as a term)?*

Response:
By 'strength', we mean sensitivity to GMT change. We will change the wording here to 'sensitivity'.
By 'simplified', we certainly did not mean to imply any negative connotation. What we do mean to imply is that we do not apply additional terms to the pattern equation like Frieler et al, 2012, and Herger et al, 2015, or adjust epoch length like Barnes & Barnes 2015.
Most studies on pattern scaling use the term "simple" to describe the standard, no additional predictors, terms, or constraints, as has been used in the Tebaldi and Arblaster, 2014, paper.

Revision:
In this manuscript, we use a simplified approach for each method to assess the differences in pattern sensitivity and skill between each method's generated pattern.

---

13. *"potential mitigation" why potential?*

Response:
Page 3, line 22, we agree that 'potential' is not needed here. It will be removed.

---

14. "*Because models varied in spatial resolution, when appropriate, the models were first regridded to T85 resolution" what does "when appropriate" refer to, here? When was not appropriate?*

Response:
Page 4, line 16: We did not regrid the models when constructing certain figures (Figure 1-3 and 5; Tables 2 & 3; S3), as regridding wasn't necessary. For spatial analysis, regridding was necessary.
Additionally, in response to another reviewer, all models were regridded to the lowest model resolution of the ensemble. We have edited and expanded this section in the text.

Revision:
All model output were regridded to lowest spatial resolution of the ensemble prior to calculating ensemble mean or median. This was done for averaging purposes, as each model had a different spatial resolution. Regridding to the lowest resolution of the multi-model ensemble was necessary as regridding to the highest resolution of the multi-model ensemble would lead to errors if used for purposes of impact assessments. A brief analysis (not shown) was done to examine the statistical difference in the multi-model ensemble mean patterns between the output of the highest and lowest regridding scheme. Statistically significant difference existed only at high latitudes for limited portions of the Arctic and Antarctic.

---

15. *The discussion of the biases in the reanalysis affecting pattern scaling methodologies is completely unintelligible to me. I am not aware of any methods of pattern scaling that uses reanalysis or other observational data.*

Response:

We have addressed this concern in response to Question/Comment #4. Again, reanalysis data was not used in construction of patterns. Other reviewers have had similar concerns, and we have addressed those questions as well.

---

16. *The discussion of different sectors using different periods of reference is again difficult to understand in this context.*

Response:
Page 4, line 30: This line is in reference to page 2, lines 26-32, and the various time periods used as a 'reference' or 'baseline' period. The paragraph on Page 4 was removed, as it was not suited to the data analysis section.

---

17. *The idea that pattern scaling should use an ensemble of models is actually open for discussion: pattern scaling was essentially developed for single models, and work in the literature has shown how high the variability is of patterns across models, so I am not sure that taking this view would satisfy any practitioner of pattern scaling. After all, the most popular pattern scaling tool, MAGICC-SCENGEN provides a library of patterns, one for each model.*

Response:
SCENGEN is a program that takes in global emission and temperature data from MAGICC, and uses the spatial climate pattern derived from CMIP3 models to produce a 2-D data product that specifies the amount of change (global or regional) projected at a specific time in the future [Wigley, 2008]. They use the "delta" method exclusively. At this point, we would like to reiterate that we are merely attempting to compare the mean patterns from a selection of models. This is not unique as many studies have used a multi-model ensemble to examine patterns. When we use these patterns in an experiment with Hector, our SCM, we will not be using the multi-model ensemble average/median.

---

18. *"we assume the ensemble variance for each pattern is the same" what does this mean? Couldn't you compute the actual ensemble variance for each pattern and use that in your statistical significance testing?*

Response:
We use the Student's t-test probability distribution which uses population variance and mean to test for a significant difference between two populations. A requirement of the Student's t-test is that the two populations being compared have equal variance. If they are not the same, then one must use Welch's t-test (or some other test with relaxed assumptions). In some analysis we did (not shown), the results for the returned probability from the two different tests didn't matter, so we went with the more familiar Student's t-test, which requires that the ensemble variance for each pattern is the same. We would like to keep in the above mentioned phrasing because it is necessary to define in the calculation of the static.

---

19. *Here and in other places a big emphasis is made on the assumption that the trend has to be the same and I fail to understand why that is the case. The delta method has been applied across scenarios, it has been shown that different scenarios, as long as they are mostly driven by increases in GHGs, produce very similar patterns, but the trends in different scenarios are*

*far from the same. So it would not seem to be a requirement, and in fact it cannot be, given that the method is developed to produce new scenarios, most of which will differ in trend.*

Response:
We agree, and we do state this explicitly in this section on page 6, line 13/14. The assumption of GMT change and local change being independent of trend is important when considering the length of the epoch. The idea is that a longer epoch will increase the signal-to-noise ratio, by decreasing variance, but may also decrease the signal. This idea is first explored by Mitchell, 2003, and later by Barnes and Barnes, 2015. Because we do not explore the issue of epoch length, we have edited this paragraph and moved a portion of it to the methods section where we justify our use of the 30-yr epoch length.

---

20. *The discussion of variance between the two reference epochs is completely obscure. I don't actually understand what variance we are talking about. That of models? That of temperature itself? And how is that relevant here?*

Response:
We agree that the discussion of variance was not clear and not well represented by the figures. The variance between models and the variance of temperature itself are explored by examining the t-test results and in Supplementary figures.

---

21."*When using the lower forcing scenario, it is not necessarily implied that there is less future variability" this is a sentence whose meaning in this context is not easy to understand.*

Response:
This sentence was not necessary, and has been removed from the manuscript.

---

22. *Is that 0.8 on line 25, page 7 suppose to be 0.08?*

Response:
Yes, thank you.

---

23. *The discussion of how lower forcing scenarios produce less change in GMT and that explains less variance explained may be justifiable is you discuss the role of internal variability, but you don't.*

Response:
We agree that variance had not been properly evaluated. We have added an additional figure in the main text as well as a figure in the supplementary material to further address this issue. Additionally, the idea of signal to noise ratio has been examined briefly to discuss the role of variance in pattern creation.

---

24. "*This is due to the fact that scenarios with stronger mitigation practices have a smaller GMT trend and the resulting local temperature sensitivity to GMT is stronger" what does this mean, and why should it be the case?*

Response:
This was poorly explained by the figures, and was perhaps more conjecture than what was shown in our results. We have added an additional figure that shows where there is a

strong difference between patterns from each scenario is also where the local/global temperature signal is significantly different.

---

25. *"Choice of scenario can affect the resulting pattern, particularly when using the regression method" I can't help wondering if this claim is the result of what I think is a mishandling of the beta coefficients.*

Response:
We acknowledge that the statement mentioned in Question/Comment #1 was incorrect and has been corrected. Our error does not invalidate the above sentence, and we believe that we can support the above statement with our results.