Mapping model agreement on future climate projections

Claudia Tebaldi,[1] Julie M. Arblaster,[2,3] and Reto Knutti[4]

Received 30 September 2011; revised 1 November 2011; accepted 1 November 2011; published 13 December 2011.

[1] Climate change projections are often based on simulations from multiple global climate models and are presented as maps with some form of stippling or measure of robustness to indicate where different models agree on the projected anthropogenically forced changes. The criteria used to determine model agreement, however, often ignore the presence of natural internal variability. We demonstrate that this leads to misleading presentations of the degree of model consensus on the sign and magnitude of the change if the ratio of the signal from the externally forced change to internal variability is low. We present a simple alternative method of depicting multimodel projections which clearly separates lack of climate change signal from lack of model agreement by assessing the degree of consensus on the significance of the change as well as the sign of the change. Our results demonstrate that the common interpretation of lack of model agreement in precipitation projections is largely an artifact of the large noise from climate variability masking the signal, an issue exacerbated by performing analyses at the grid point scale. We argue that separating more clearly the case of lack of agreement from the case of lack of signal will add valuable information for stake-holders' decision making, since adaptation measures required in the two cases are potentially very different. **Citation:** Tebaldi, C., J. M. Arblaster, and R. Knutti (2011), Mapping model agreement on future climate projections, *Geophys. Res. Lett.*, *38*, L23701, doi:10.1029/2011GL049863.

1. Introduction

[2] Different global climate models produce different outcomes for future climate change even under the same future pathway of greenhouse gas concentrations. Methods are being developed that try to synthesize different projections in the now paradigmatic multimodel approach [*Knutti et al.*, 2010; *Meehl et al.*, 2007a; *Smith et al.*, 2009; *Tebaldi and Knutti*, 2007; *Tebaldi et al.*, 2006], but in many cases only simple criteria are used to quantify and display agreement of the projected anthropogenic changes, e.g. the ratio between the spread across models (measured as one or two standard deviations) compared to the multimodel mean response [*Deser et al.*, 2011; *Meehl et al.*, 2007b, Figure 10.9], or the number of models agreeing on the sign of change, adopted in the *Intergovernmental Panel on Climate Change*'s [2007] Figure SPM.7 for precipitation (SPM.7 from now on). The idea is that if multiple models, based on different but plausible assumptions, simplifications and parameterizations, agree on a result, we have higher confidence than if the result is based on a single model, or if models disagree on the result. A more in-depth discussion of this point is given by *Räisänen* [2007] and *Schaller et al.* [2011].

[3] As pointed out by recent studies [*Deser et al.*, 2011; *Hawkins and Sutton*, 2009, 2011], a major source of uncertainty besides model spread is internal natural variability of the system. This becomes increasingly relevant as attention is focused on short term projections and predictions and decadal predictability experiments are performed [*Meehl et al.*, 2009], and as interest is focused on regional details of future changes. At these shorter timescales and smaller spatial scales the climate change signal decreases relative to the internally generated noise of the climate system [*Mahlstein et al.*, 2011]. Temperature projections benefit from a high signal-to-noise ratio even for small spatial scales and short term horizons, while precipitation change has the opposite characteristic. Attribution studies also confirm this dichotomy. The signal of an externally forced temperature change has already emerged from the noise generated by natural variability in all continents [*Stott*, 2003] while changes outside of natural variability in precipitation have been detected only for a zonal mean pattern over the whole globe [*Zhang et al.*, 2007]. Internal variability dominates at the grid point scale for precipitation projections over the next few decades [*Deser et al.*, 2011; *Hawkins and Sutton*, 2011], so for short term projections and variables with low signal-to-noise ratios, simple criteria for model agreement of the forced change that do not take into account the effect of natural variability are prone to misinterpretation when they equate lack of model consensus with lack of information.

[4] Representations of future projections in temperature and precipitation as global maps, of the type found in SPM.7 (see, e.g., Figure 1c) may lead to such misinterpretations of climate change projections, and we propose a new method addressing this limitation. Of particular concern are swaths of white that cover large regions (where the model consensus on the sign of the change is less than 66%) and the sparseness of the stippling in SPM.7 (where the consensus is less than 90%). The typical interpretation of the white areas is that projections for precipitation are inconsistent between different models [*Anderson et al.*, 2009]. But as pointed out recently, the lack of robust trends is partly attributable to a low signal-to-noise ratio, rather than inconsistent model responses [*Schaller et al.*, 2011; *Power et al.*, 2011]. There is a fundamental difference between lack of signal (i.e., lack of detection of a significant response to external forcing) versus lack of agreement in the signal, i.e. between regions where the change is not statistically significant and regions where different models produce significant changes of opposite sign (disagreement over the magnitude of change

[1]Climate Central, Princeton, New Jersey, USA.
[2]Center for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Victoria, Australia.
[3]Climate and Global Dynamics, National Center for Atmospheric Research, Boulder, Colorado, USA.
[4]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland.

0094-8276/11/2011GL049863

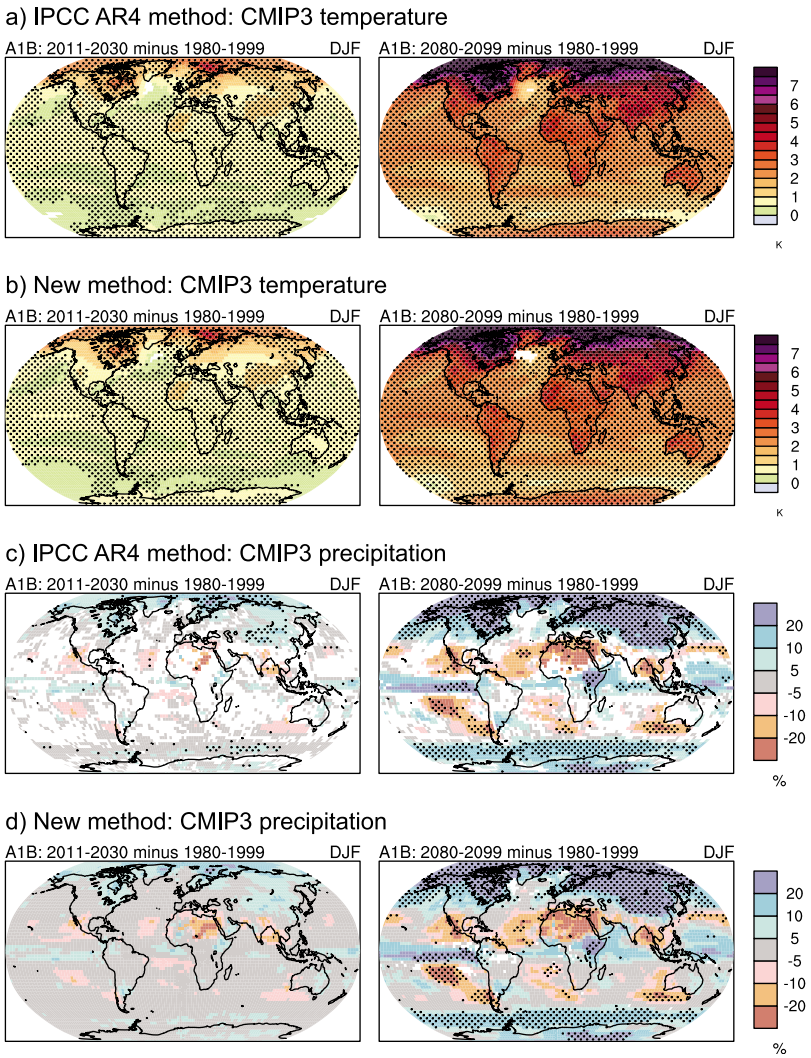

Figure 1. (left) Early (2020) and (right) late (2090) century projections of December to February (a and b) surface temperature and (c and d) precipitation change from the CMIP3 models. Our new method (Figures 1b and 1d) is compared with the method that produced the AR4 SPM figures (Figures 1a and 1c).

is arguably less troubling for such a summary representation, but would add another dimension to the problem). Arguably, decisions related to adaptation considering these two types of climate projections would come to very different conclusions: in the former case (lack of signal), stake-holders may use the information contained in the (observed) historical variability of precipitation, coupled with the usually more robust temperature projections, to devise adaptation strategies, whereas in the latter case (lack of agreement) stake-holders are left worrying about a future that promises significantly changed conditions, but in an uncertain direction, feeling paralyzed by the lack of actionable information [*Jones and Boer*, 2005; *Moser*, 2011]. Simple tests of agreement in sign, e.g., in SPM.7, confound the two issues: Simply put, in the case of negligible change that is still within the noise of the system, models have a good chance of not agreeing on the sign of change but still agree on the negligible nature of that change. This is very different from the case where models produce contradicting predictions over the direction of a significant change that the climate system will experience in reaction to increased greenhouse gases.

[5] Here we propose a simple method of analysis and graphical representation that will clearly delineate the difference between these cases. We use examples of both temperature and precipitation since the two variables are significantly different with regard to signal-to-noise ratios and degree of model agreement over the sign, size and significance of the forced change. The method should be applicable to other variables as well, which may have characteristics more similar to one end of the spectrum (e.g. temperature) or the other (e.g. precipitation), as far as the role of natural variability is concerned.

[6] Our goal in this paper is to describe a method that succinctly, comprehensively and transparently depicts the results of multimodel projections in accordance with our discussion of the shortcoming of SPM-style projection maps. We do not address the reasons behind the models' future behavior, nor issues of model validation, biases and differential weighting of the multimodel ensemble members. We are simply proposing a presentation that we believe enhances interpretation and understanding of future projections from multimodel

ensembles, when low signal-to-noise or model disagreement play a role, individually or in concert.

2. Method

[7] SPM.7 represented by colors the value of the multimodel averages and by stippling the areas where at least 90% of the models agreed on the sign of the change. When less than 66% of the models agreed in sign the map was left white, to indicate lack of agreement and therefore lack of any robust information about the direction of future change.

[8] Our method explicitly considers statistical significance in the choice of coloring or not, and stippling or not. Differently from SPM.7, therefore, we distinguish the case where models do not agree in sign but are still within the boundaries of natural variability - in which case we argue that information is available, and we still use colors to represent the multimodel mean – from the case where models do not agree and simulate a significant change – in which case we argue that we truly have conflicting information, originating from the different models different responses to forcings – and we leave the corresponding areas white. There will be areas where the emergence of the signal from the noise will happen consistently across the multimodel ensemble (a majority of models will agree on significance and sign). For these areas we will use color to indicate the multimodel mean and stippling to indicate agreement in the significance and the sign.

[9] The method thus uses the following steps, grid point by grid point (note that our results will be dependent on the resolution of model output, and on the level of regional aggregation that is performed before analyzing the significance of the changes): 1) Test for significant change in each of the models individually with a t-test comparing the mean of the reference and the future period, 2a) if less than X = 50% of the models show a significant change then show the multimodel mean change in color, 2b) if more than 50% of the models show significant change then test for agreement in sign by the following criteria, 3a) if less than Y = 80% of the significant models agree on the sign then show the grid point as white, 3b) if more than 80% agree on the sign show it in color with stippling. The X and Y percentages are of course a subjective choice. They could be chosen differently depending on the desired level of confidence. Also note that consistency in the sign of the forced signal is considered here, but other criteria could be devised to consider agreement in magnitude. The conceptual idea would be similar in all cases.

[10] One could take a more formal approach to the choice of X and Y considering that we can regard the behavior of each model (significant or not, agreeing in sign or not) as the realization of a binary variable having – under the null hypothesis – 50% chance of turning out 0 or 1. Under this model, with p = 0.5, we can compute the expected number of successes for N trials, N being the number of models considered, which equals p·N, and the variance of the distribution of successes, equal to $N \cdot p \cdot (1 - p)$. We can then choose a range that covers 95% of the probability for the variable "number of successes" under the hypothesis of random and independent trials (leaving the issue of characterizing model dependence for other discussions) and choose X and Y accordingly, thus protecting ourselves from random occurrences of disagreement. We are not, in this paper, especially focused on the values of X and Y. In particular, we chose not to replicate the IPCC choices of 66% and 90% in order not to draw special attention to these quantities, which coincide with what IPCC uses as boundaries for a probabilistic statement to signify a likely (>66%) or very likely (>90%) outcome [*Mastrandrea et al.*, 2010]. We are here considering a fraction of models from an ensemble of opportunity and we want to explicitly separate our choices of X and Y from more formal assessment of confidence or likelihood, which would necessitate further considerations (e.g., of model dependencies, sampling, model performance, and common structural errors) than simple empirical frequencies from a multimodel ensemble.

[11] Other more sophisticated methods to quantify internal variability (e.g. using control runs or multiple ensemble members for each model) are possible, but again our concept is generic. The criteria used here are deliberately kept simple and transparent, and only one transient simulation from each model is required. The proposed measures do not consider model dependence [*Masson and Knutti*, 2011a; *Pirtle et al.*, 2010]. The significance of the signal and model agreement on it also depends on the spatial scale [*Hawkins and Sutton*, 2011; *Masson and Knutti*, 2011b], and model agreement has been shown to be better if regions with similar base climate and change are carefully chosen [*Mahlstein and Knutti*, 2010].

3. Results

[12] The results of the original method used in SPM.7 and the new method are shown in Figure 1 for short term and long term projections and for both temperature and precipitation. Results are shown for December to February for illustration; results for June to August are given in the auxiliary material.[1] 21 models from the CMIP3 archive [*Meehl et al.*, 2007a] are used.

[13] For temperature, changes soon are significant and models agree on the sign (Figure 1a). The two methods produce results that are almost identical (Figures 1a and 1b). For precipitation, using the IPCC method and looking out to 2020 (Figure 1c), however, the map is mostly white, but in fact models agree that the signal is just small and has not emerged from noise (as our new method depicted in Figure 1d clearly shows). For the new method both the number of white grid points and those with stippling increase with time, as expected as the signal emerges, but the overall pattern of change is similar for both time periods.

[14] An interesting test is to apply the two methods to an initial condition ensemble of a single model (in this case 8 members from an initial condition ensemble with CCSM3). Figure 2 shows that even in this case the IPCC method produces large white areas in a picture like SPM.7. This makes no sense, as in this case model uncertainty is absent altogether and, by construction, there must be no inconsistency of model response among the different simulations. The new method shows clearly that the early decades have no significant signal. Towards the end of the century, some areas start to show significant signals, but there are no white areas indicating inconsistency since all members come from the same model (Figure 2b, right).

[1] Auxiliary materials are available in the HTML. doi:10.1029/2011GL049863.

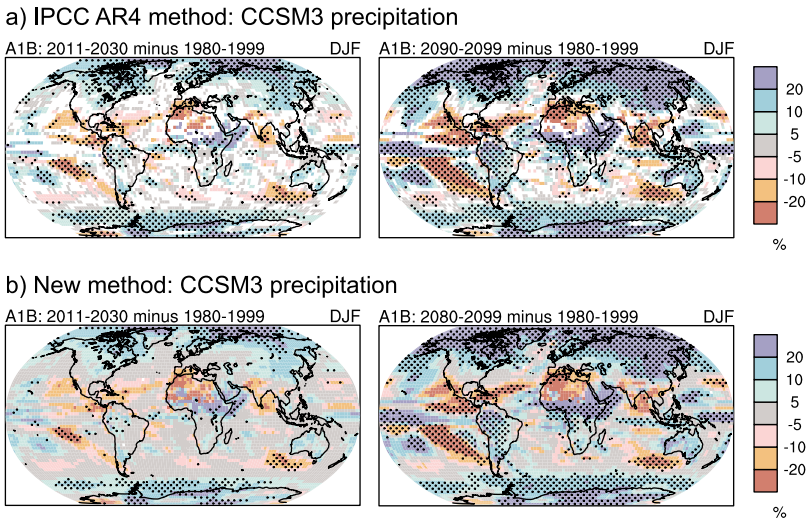

Figure 2. (left) Early (2020) and (right) late (2090) century projections of December to February precipitation change from eight initial condition ensemble members of the NCAR CCSM3, for (a) the AR4 SPM and (b) our new method.

[15] In the methodology proposed in this paper we are testing for significance within each model by using the variance estimated from within each of the model runs. We therefore want to test that our method is robust when using another measure of natural variability which is often considered, i.e., the different realizations that are available from an individual model's ensemble members. Because only a small subset of models have at least three ensemble members available under a given emission scenario, we artificially construct three multimodel ensembles that sample the internal variability of each model as follows. We use all the models that performed runs under the three SRES scenarios prescribed under CMIP3 (B1, A1B and A2), treating the three individual scenario runs for each model as a surrogate of an initial condition ensemble. In order to make the signal of climate change comparable across these three "members" and use them to only span a range of variations due to natural variability and model uncertainty (but not scenarios), we select the 20yr period around the time when the multimodel mean global average surface temperature under each scenario reaches 1°C above the reference period. This procedure thus produces three ensembles including the same 14 models producing climate change of similar magnitude, but made of runs whose individual behavior spans three realizations of natural variability. We indeed find that our new method, similarly to the SPM method, produces maps that are very similar when comparing the three ensembles, confirming that our method maintains the desired quality of not being strongly dependent on the particular sampling of natural variability (the specific run included for each model when more than one run is available), at least for an ensemble of the typical CMIP3 size (see Figures S3 and S4 in the auxiliary material).

4. Discussion

[16] We propose a succinct and intuitive way to display changes and agreement among models in a multimodel ensemble that clearly separates lack of signal from lack of information due to model disagreement. We thus categorize three levels of multimodel agreement: 1) the majority of models agree that future changes will be statistically significant and of the same sign 2) the majority of models show significant change but in opposite directions and 3) most of the models show no significant change. The basic idea is that testing for model agreement is only meaningful if the models are producing significant changes, i.e., changes outside of internal variability. Apart from this conceptual advance, a few conclusions are worth highlighting. First, in contrast to popular belief, model agreement of future precipitation change is greater than currently thought. Only few places in the world show significant changes of opposite sign in different models. Second, despite a clear anthropogenic large-scale signal, projections of precipitation at the grid point scale for the next few decades are not significant for most regions. Arguing about model consistency of the sign of the signal is misplaced in this context. Third, there are large regions where we are quite confident that anthropogenically forced changes are likely to be small in the next few decades, information that is no doubt useful for adaptation. Obviously, the details of our analysis depend on the spatial resolution adopted, which we chose here as T42 (about 250 by 250 km in grid box size), the same resolution that was adopted to process and display multimodel results in the last IPCC report's SPM.

[17] This paper focuses on a methodology that is as simple and transparent as possible. We do not address issues of dependency among models, model evaluation or weighting, or more sophisticated approaches to characterizing significant change at the grid point or the field level, neither do we address explicitly the problem of multiple comparisons when testing a field grid point by grid point (except to say that an application of the False Discovery Ratio methodology [*Ventura et al.*, 2004] did not change our results in any appreciable way). We hope that researchers can take this generic approach and fill in the steps having to do with the definition of significance and the definition of agreement in the way that best suit their analysis' foci and goals.

[18] **Acknowledgments.** We would like to thank Joel B. Smith and an anonymous reviewer for their careful reading and thoughtful feedback. We also thank Scott Power for discussions about the general topic of this article and his work. Portions of this study were supported by the Office of Science, Biological and Environmental Research, U.S. Department of Energy, Cooperative Agreement DE-FC02-97ER62402, DOE's Office of Biological and Environmental Research grant DE-SC0004956, and the National Science Foundation. Claudia Tebaldi is grateful to NCAR/CGD for hosting her. We acknowledge the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multimodel dataset. Support of this dataset is provided by the Office of Science, US DOE. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

[19] The editor thanks Joel Smith and an anonymous reviewer.

References

Anderson, B. T., C. Reifen, and R. Toumi (2009), Consistency in global climate change model predictions of regional precipitation trends, *Earth Interact.*, *13*, Paper 9, 1–23, doi:10.1175/2009EI273.1.

Deser, C., A. Phillips, V. Bourdette, and H. Teng (2011), Uncertainty in climate change projections: the role of internal variability, *Clim. Dyn.*, doi:10.1007/s00382-00010-00977-x, in press.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*(8), 1095–1107, doi:10.1175/2009BAMS2607.1091.

Hawkins, E., and R. Sutton (2011), The potential to narrow uncertainty in projections of regional precipitation change, *Clim. Dyn.*, *37*(1–2), 407–418, doi:10.1007/s00382-010-0810-6.

Intergovernmental Panel on Climate Change (2007), Summary for policymakers, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 1–18, Cambridge Univ. Press, Cambridge, U. K.

Jones, R., and R. Boer (2005), Assessing current climate risks, in *Adaptation Policy Frameworks for Climate Change: Developing Strategies, Policies and Measures*, edited by B. Lim et al., pp. 91–117, Cambridge Univ. Press, Cambridge, U. K.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl (2010), Challenges in combining projections from multiple climate models, *J. Clim.*, *23*(10), 2739–2758, doi:10.1175/2009JCLI3361.1.

Mahlstein, I., and R. Knutti (2010), Regional climate change patterns identified by cluster analysis, *Clim. Dyn.*, *35*(4), 587–600, doi:10.1007/s00382-009-0654-0.

Mahlstein, I., R. Knutti, S. Solomon, and R. Portmann (2011), Early onset of significant local warming in low latitude countries, *Environ. Res. Lett.*, *6*(3), 034009, doi:10.1088/1748-9326/6/3/034009.

Masson, D., and R. Knutti (2011a), Climate model genealogy, *Geophys. Res. Lett.*, *38*, L08703, doi:10.1029/2011GL046864.

Masson, D., and R. Knutti (2011b), Spatial-scale dependence of climate model performance in the CMIP3 ensemble, *J. Clim.*, *24*(11), 2680–2692, doi:10.1175/2011JCLI3513.1.

Mastrandrea, M., et al. (2010), Guidance note for lead authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties, report, 5 pp., Intergov. Panel on Clim. Change, Geneva, Switzerland.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor (2007a), The WCRP CMIP3 multimodel dataset—A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*(9), 1383–1394, doi:10.1175/BAMS-88-9-1383.

Meehl, G. A., et al. (2007b), Global climate projections, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 747–846, Cambridge Univ. Press, Cambridge, U. K.

Meehl, G. A., et al. (2009), Decadal prediction: Can it be skillful?, *Bull. Am. Meteorol. Soc.*, *90*(10), 1467–1485, doi:10.1175/2009BAMS2778.1.

Moser, S. (2011), The contextual importance of uncertainty in climate-sensitive decision-making: Toward an integrative decision-centered screening tool, in *Climate Change in the Great Lakes Region: Navigating an Uncertain Future*, edited by T. D. D. Bidwell, pp. 179–212, Mich. State Univ. Press, East Lansing.

Pirtle, Z., R. Meyer, and A. Hamilton (2010), What does it mean when climate models agree? A case for assessing independence among general circulation models, *Environ. Sci. Policy*, *13*(5), 351–361, doi:10.1016/j.envsci.2010.04.004.

Power, S., et al. (2011), Consensus on 21st century rainfall projections in climate models more widespread than previously thought, *J. Clim.*, in press.

Räisänen, J. (2007), How reliable are climate models?, *Tellus, Ser. A*, *59*(1), 2–29.

Schaller, N., I. Mahlstein, J. Cermak, and R. Knutti (2011), Analyzing precipitation projections: A comparison of different approaches to climate model evaluation, *J. Geophys. Res.*, *116*, D10118, doi:10.1029/2010JD014963.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (2009), Bayesian modeling of uncertainty in ensembles of climate models, *J. Am. Stat. Assoc.*, *104*(485), 97–116, doi:10.1198/jasa.2009.0007.

Stott, P. A. (2003), Attribution of regional-scale temperature changes to anthropogenic and natural causes, *Geophys. Res. Lett.*, *30*(14), 1728, doi:10.1029/2003GL017324.

Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philos. Trans. R. Soc. A*, *365*(1857), 2053–2075.

Tebaldi, C., K. Hayhoe, J. M. Arblaster, and G. A. Meehl (2006), Going to the extremes, *Clim. Change*, *79*(3–4), 185–211, doi:10.1007/s10584-006-9051-4.

Ventura, V., C. J. Paciorek, and J. S. Risbey (2004), Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data, *J. Clim.*, *17*(22), 4343–4356, doi:10.1175/3199.1.

Zhang, X. B., F. W. Zwiers, G. C. Hegerl, F. H. Lambert, N. P. Gillett, S. Solomon, P. A. Stott, and T. Nozawa (2007), Detection of human influence on twentieth-century precipitation trends, *Nature*, *448*(7152), 461–465, doi:10.1038/nature06025.

J. M. Arblaster, Bureau of Meteorology, BMRC, GPO Box 1289, Melbourne, Victoria 3001, Australia.

R. Knutti, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstr. 16, CH-8092 Zurich, Switzerland.

C. Tebaldi, Climate Central, c/o National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000, USA. (claudia.tebaldi@gmail.com)