Emerging selection bias in large-scale climate change simulations

Kyle L. Swanson[1]

Received 3 April 2013; revised 10 May 2013; accepted 13 May 2013; published 18 June 2013.

[1] Climate change simulations are the output of enormously complicated models containing resolved and parameterized physical processes ranging in scale from microns to the size of the Earth itself. Given this complexity, the application of subjective criteria in model development is inevitable. Here we show one danger of the use of such criteria in the construction of these simulations, namely the apparent emergence of a selection bias between generations of these simulations. Earlier generation ensembles of model simulations are shown to possess sufficient diversity to capture recent observed shifts in both the mean surface air temperature as well as the frequency of extreme monthly mean temperature events due to climate warming. However, current generation ensembles of model simulations are statistically inconsistent with these observed shifts, despite a marked reduction in the spread among ensemble members that by itself suggests convergence towards some common solution. This convergence indicates the possibility of a selection bias based upon warming rate. It is hypothesized that this bias is driven by the desire to more accurately capture the observed recent acceleration of warming in the Arctic and corresponding decline in Arctic sea ice. However, this convergence is difficult to justify given the significant and widening discrepancy between the modeled and observed warming rates outside of the Arctic. **Citation:** Swanson, K. L. (2013), Emerging selection bias in large-scale climate change simulations, *Geophys. Res. Lett.*, *40*, 3184–3188, doi:10.1002/grl.50562.

1. Introduction

[2] Selection biases in information processing occur when expectations affect behavior in a manner that makes those expectations come true [*Nickerson*, 1998; *Poletiek*, 2001]. The Nobel Prize-winning physicist Richard Feynman referred to one particularly notable example in the history of physics that occurred following Robert A. Millikan's original measurement of the charge of the electron [*Feynman and Leighton* 1985]. Millikan's original measurement was slightly erroneous due to the use of an incorrect value of the viscosity of air. In the decades following Millikan's work and his subsequent Nobel Prize, other investigators empirically measured the electron charge. The values they obtained show a curious trend, creeping further and further away from Millikan's canonical value until finally settling down at the modern figure. To quote Feynman: *When they got a number that was too high above Millikan's, they thought something must be wrong–and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that* [*Feynman and Leighton* 1985 pg. 342]. Selection bias, involving a choice of which observations were kept based upon a prior canonical but erroneous experimental result, inhibited progress in scientific endeavor.

[3] Here we suggest the possibility that a selection bias based upon warming rate is emerging in the enterprise of large-scale climate change simulation. Instead of involving a choice of whether to keep or discard an observation based upon a prior expectation, we hypothesize that this selection bias involves the 'survival' of climate models from generation to generation, based upon their warming rate. One plausible explanation suggests this bias originates in the desirable goal to more accurately capture the most spectacular observed manifestation of recent warming, namely the ongoing Arctic amplification of warming and accompanying collapse in Arctic sea ice. However, fidelity to the observed Arctic warming is not equivalent to fidelity in capturing the overall pattern of climate warming. As a result, the current generation (CMIP5) model ensemble mean performs worse at capturing the observed latitudinal structure of warming than the earlier generation (CMIP3) model ensemble. This is despite a marked reduction in the interensemble spread going from CMIP3 to CMIP5, which by itself indicates higher confidence in the consensus solution. In other words, CMIP5 simulations viewed in aggregate appear to provide a more precise, but less accurate picture of actual climate warming compared to CMIP3.

[4] We raise this issue in the context of two simple but related questions. First, how well do climate simulations capture the latitudinal pattern of warming when the most recent decade (2002–2011) is compared with the mean surface air temperature of the 1979–2001 period marked by the advent of intensive satellite observation of Earth's surface? This question is of intrinsic interest, as it captures the leading order essence of climate warming issue itself. A reasonable expectation is that climate simulations should exhibit increased fidelity to the observed climate change signal as they develop from generation to generation. Barring increased fidelity, they should at least retain the statistical diversity necessary to encapsulate the observed warming.

[5] The second question concerns higher moments of the spatial pattern of climate change and may be phrased as follows: Over the most recent decade (2002–2011), has the frequency of local anomalously warm (or cold) months relative to the entire 1979–2011 period changed? This question is of societal interest, as there has been recent speculation that increased incidence of extreme monthly temperature anomalies is a harbinger of looming global

[1] Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA.

Corresponding author: K. L. Swanson, Atmospheric Sciences Group, Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA. (kswanson@uwm.edu)

0094-8276/13/10.1002/grl.50562

warming impacts [*Hansen et al.*, 2012; *Trenberth and Fasullo*, 2012]. We seek to examine these questions by studying how climate simulation behavior has changed from one model generation to the next, examining how aggregate simulation behavior evolves as models become more advanced and computational power increases.

2. Methods and Analysis

[6] We seek to compare the observed mean warming of the near-surface air temperature as well as the frequency of anomalously warm and cold months against projections made by climate change simulations. Over the past decade, many institutions have made the output of their climate change simulation models available through the Coupled Model Intercomparison Program (CMIP) projects under the auspices of the World Climate Research Program (WCRP), with each project spanning a generation of these models. This provides us with unique insight into how the models that generate these climate change simulations are themselves evolving with time. We specifically consider here two major recent projects; the CMIP3 project that captured the state of models at the year 2005, and the CMIP5 project that captures the current state of such models (2012).

[7] Data used from the CMIP3 project are the near-surface air temperature ('tas') monthly mean fields 1979–2011. These fields were generated from simulations driven by historical forcing until the early 2000s and continued into the 21st century using the SRES A1B 'business as usual' forcing scenario [*Meehl et al.*, 2007]. Data were downloaded from the CMIP3 data portal maintained by the Program for Climate Model Diagnostics and Intercomparison (PCMDI; http://cmip-pcmdi.llnl.gov/). Data from the CMIP5 project are the near-surface air temperature ('tas') monthly mean fields 1979–2011. These fields were generated from simulations driven by historical forcing until the end of 2005 and continued into the 21st century using the RCP4.5 (medium CO2 emission mitigation) forcing scenario [*Taylor et al.*, 2012]. Data fields were downloaded from the CMIP5 data portal, also maintained by PCMDI (http://cmip-pcmdi.llnl.gov/).

[8] These climate model simulations are compared against near-surface air temperature ('t2m') reanalysis fields originating from the European Center for Medium Range Forecasts (ECMWF) Intermediate Reanalysis Project [ERA-Interim; *Dee et al.*, 2011]. This reanalysis blends in situ and satellite observations with short-range numerical weather prediction model simulations using an advanced assimilation scheme to provide a best guess for the atmospheric state. The ERA-Interim reanalysis is the current state of the art. This data was downloaded from the ECMWF data portal (http://data-portal.ecmwf.int/). Further comparison is made against the HadCRUT4 observed surface air temperature fields [*Morice et al.*, 2012].

[9] Figure 1A shows the change in near-surface air temperatures for the 2002–2011 decade relative to the 1979–2001 mean as projected by the 52 CMIP3 model simulations (23 unique models) for the tropics (equatorward of 30° latitude) and extratropics (poleward of 30°), along with the analogous change for the ERA-Interim reanalysis and HadCRUT4 observations. As expected, both the HadCRUT4 and reanalysis surface temperatures are warmer in both the tropics and extratropics in the most recent decade compared to their respective 1979–2001 means. The model simulations in general share this tendency, although in aggregate they overpredict the warming in both the tropics and extratropics. This in itself is not problematic; models are inherently imperfect, and such imperfections do not by themselves limit the usefulness of these model projections provided they are properly treated [*Smith*, 2002; *Raisanen*, 2007]. What is significant is that the HadCRUT4 and reanalysis warming lie well within the spread of the model simulations.

[10] Figure 1B shows the change in near-surface air temperatures for the 2002–2011 decade relative to the 1979–2001 mean as projected by the 92 model simulations (38 unique models) for the CMIP5 project, again with the analogous change for the ERA-Interim reanalysis and HadCRUT4 observations for comparison. Curiously, simulation analogues for the observed warming have largely disappeared in the CMIP5 project. The HadCRUT4 and reanalysis warming lie on the fringes of the model envelope, roughly 2 standard deviations (internally calculated from the intersimulation spread) removed from the model simulation ensemble mean. Curiously, the CMIP5 simulations appear to be approaching a consensus, as the intersimulation standard deviation is 25% smaller among the CMIP5 project simulations than among the CMIP3 project simulations (Table 1). However, this consensus appears to explicitly exclude the observed warming.

[11] The latitudinal structure of the warming shown in Figures 1C and 1D provides insight into the unusual behavior exhibited by the CMIP5 ensemble. In the CMIP3 ensemble, the largest deviation between observed and simulated warmings is in the Arctic, where the observed warmings are roughly 1°C larger than the CMIP3 simulation ensemble mean. The CMIP5 ensemble successfully reduces this deviation in the Arctic (Figure 1D), with differences in the warming pattern between the CMIP5 and CMIP3 ensemble means outside of the Arctic consistent with diffusion of the enhanced CMIP5 warming in the Arctic into the Northern Hemisphere midlatitudes. However, the enhanced CMIP5 ensemble mean Arctic warming unveils offsetting errors in the CMIP3 ensemble mean warming (not enough warming in the Arctic, too much warming almost everywhere else), leading to the poorer overall CMIP5 ensemble mean consistency with the observed warming relative to CMIP3.

[12] This description provides a reasonable explanation for why the CMIP5 ensemble mean performs poorly relative to CMIP3. However, the issue of the reduction in the CMIP5 simulation spread still remains. One way to approach this problem is to ask what subset of the CMIP3 ensemble has statistics most like the CMIP5 ensemble. To this end, consider a subensemble comprised of those CMIP3 simulations that warm more than the ensemble median CMIP3 simulation (hereafter CMIP3+). Curiously, the statistics of this CMIP3+ subensemble are indistinguishable from those of the CMIP5 ensemble using Student's *T*-test (Table 1; $p \simeq 0.15$ for both tropics and extratropics). This contrasts with the behavior of the entire CMIP3 ensemble, which differs from the CMIP5 ensemble in a statistically significant fashion in both the tropics and extratropics ($T > 3.25$; $p < .002$). This indicates the possibility of a selection bias based upon warming rate (either globally or regionally in the Arctic), with only those model configurations that warmed more aggressively 'surviving' in an appropriate sense to be included in CMIP5, while those that did not warm as aggressively were more significantly modified. This statement is of course highly speculative; the actual rationale for this convergence is likely to be more complicated.

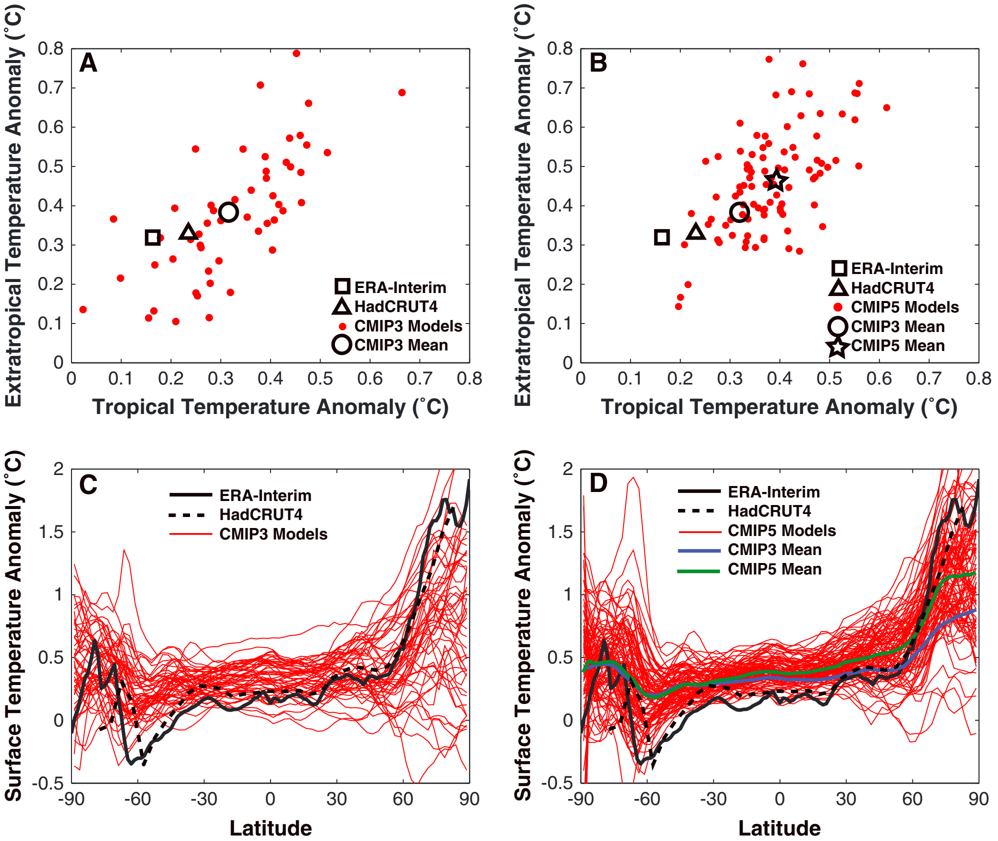

Figure 1. Changes in mean surface air temperature are the standard metric used to assess climate change. Panel A shows tropical and extratropical temperature anomalies for the CMIP3 project simulations (red dots) for the decade 2002–2011 relative to the 1979–2001 mean. Values for the ERA-Interim reanalysis and HadCRUT4 are shown for comparison. Panel B is similar, except that individual simulations are taken from the CMIP5 project. Panel C shows these surface air temperature anomalies as a function of latitude for the CMIP3 simulations (red curves), as well as for the ERA-Interim reanalysis (heavy black curve). Panel D shows the same but for the CMIP5 simulations, with the CMIP3 and CMIP5 mean simulation curves inserted for reference.

3. Anomalous Events

[13] The frequency of anomalous events provides a much stiffer test of simulation fidelity than changes to the mean. For simplicity, we define ‘anomalous’ as being among the three warmest (or coldest) out of the 33 years in the 1979–2011 period, where each month is treated separately, i.e., Octobers are only compared against other Octobers at any given location, and where we average over all months of the calendar year. These localized frequencies are then averaged spatially to facilitate comparison between models and observations. Within this context, the null expectation is that relative to the entire period, slightly less than one anomalously warm and cold month (30/33) should have occurred in the most recent decade. Since the climate has warmed over the 1979–2011 period [*Morice et al.*, 2012], we expect a higher frequency of anomalously warm months and a lower frequency of anomalously cold months during the most recent decade. With the averaging used here, a frequency of 3 months would imply that *all* anomalously warm or cold monthly events over this 33 year period at a given location occurred during the most recent decade.

[14] Figure 2A shows the frequency of anomalously warm and cold months during the most recent decade (2002–2011) relative to the entire 1979–2011 period as projected by the 52 model simulations (23 unique models) for the CMIP3 project, along with the ERA-Interim and HadCRUT4 frequencies for regions within the tropics and in the extratropics. As expected, the ERA-Interim and HadCRUT4 frequencies of extreme monthly events are consistent with warming in both the tropics and extratropics, marked by the increase in the frequency of anomalously warm months and a decrease

Table 1. Tropical and Extratropical Mean Temperatures for the Decade 2002–2011 Relative to 1979–2001 for the ERA-Interim Reanalysis, HadCRUT4, CMIP3, CMIP5, and CMIP3+ (CMIP3 above median) Climate Simulation Ensembles

	ERA-Interim	HadCRUT4	CMIP3	CMIP5	CMIP3+
Tropics	0.16 ± 0.14°C	0.23 ± 0.14°C	0.32 ± 0.12°C	0.38 ± 0.09°C	0.41 ±0.08°C
Extratropics	0.31 ± 0.14°C	0.34 ± 0.14°C	0.38 ± 0.16°C	0.46 ± 0.13°C	0.50 ± 0.12°C

Quoted uncertainties are from the intersimulation spreads in the model ensembles and consistent with a global mean temperature uncertainty of 0.1°C for global scale surface air temperature analyses (ERA-Interim and HadCRUT4) [*Arndt et al.*, 2010].

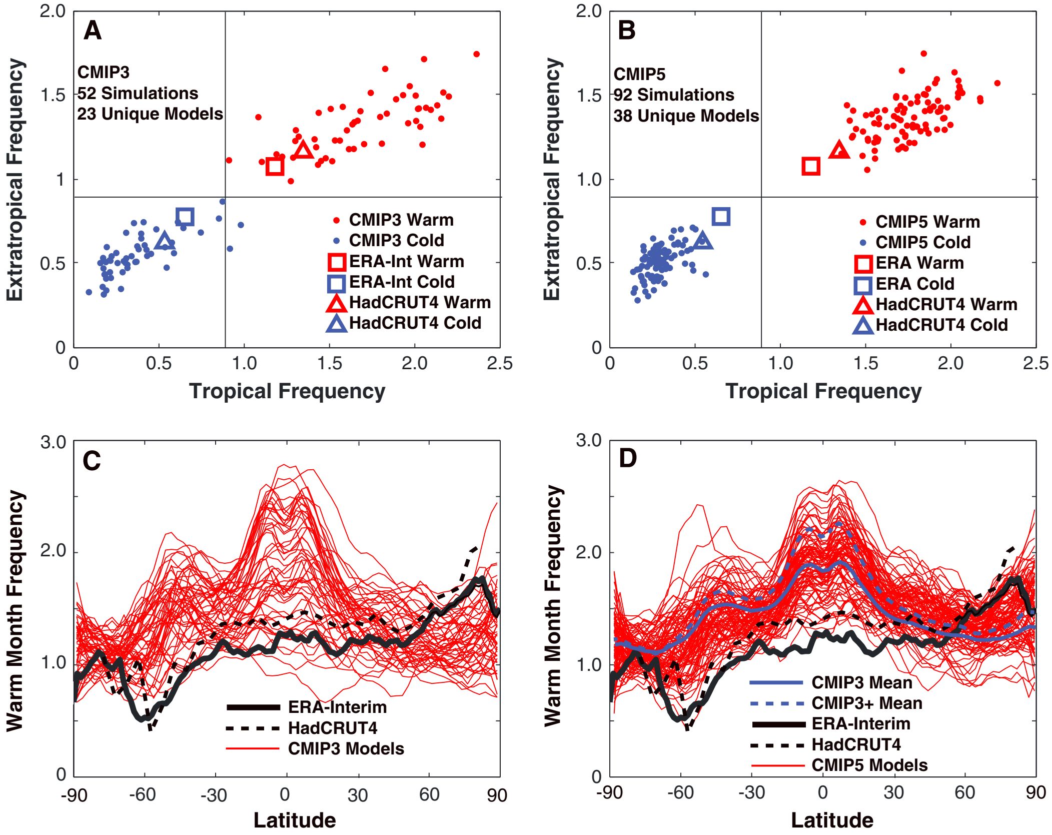

Figure 2. The probability of extreme monthly temperature events provides an objective metric to assess climate change simulations. Panel A shows the frequencies of anomalously warm (red) and cold (blue) months in the ERA-Interim reanalysis (squares) and HadCRUT4 (triangles) during the decade 2002–2011 for the extratropics (ordinate) and tropics (abscissa), relative to the 1979–2011 period of intensive atmospheric observation. Individual simulations from the CMIP3 project for the same time periods are shown as dots. Panel B is similar, except that individual simulations are taken from the CMIP5 project. Panel C shows the frequencies of anomalous warm months as a function of latitude for the CMIP3 simulations (red curves), as well as for the ERA-Interim reanalysis (heavy black curve). Panel D shows the same but for the CMIP5 simulations, with the CMIP3 and CMIP3+ ensemble mean curves inserted for reference.

in the frequency of anomalously cold months during the most recent decade relative to the entire 1979–2011 period. The model simulations share this general tendency, although in general they overpredict the frequency of anomalously warm months and underpredict the frequency of anomalously cold months relative to the observations As above, it is significant that the observed frequencies lie within the spread of the model simulations, i.e., there are simulation analogues for the observed frequencies of anomalously warm and cold months in both the tropics and extratropics.

[15] Figure 2B shows the frequencies of anomalously warm and cold months during 2002–2011 as projected by the 92 model simulations (38 unique models) for the more recent CMIP5 project, again with the ERA and HadCRUT4 frequencies for comparison. Simulation analogues for the ERA frequencies of anomalously warm and cold months in the most recent decade have disappeared in the CMIP5 project, between 3 and 4 standard deviations (internally calculated from the intersimulation spread) removed from the model simulation ensemble mean for both anomalous warm and cold monthly events. HadCRUT4 frequencies lie on the edge of the model envelope, more than 2 standard deviations removed from the model simulation mean. In spite of this, the frequency of anomalously warm and cold months in these CMIP5 simulations appears to be approaching a consensus, as the intersimulation standard deviations are roughly 50% smaller among the CMIP5 project simulations compared to the CMIP3 project simulations. However, this consensus appears to explicitly exclude the observed behavior.

[16] There are a number of complicating factors in this analysis of anomalous monthly temperature events. The most significant of these is the relative importance of the local magnitude of the climate change signal to the ‘noise’ of internal weather-related variability. High levels of weather noise would act to drive the frequency of anomalously warm and cold months towards the expectation value of slightly less than one per decade. If the CMIP5 simulations are deficient in weather-related variability, this would potentially explain the discrepancy between the observed and simulated incidences of anomalously warm and cold months. However, this appears unlikely, as it is tantamount to stating that CMIP5 simulations in aggregate have a poorer representation of weather variability than CMIP3.

[17] To gain insight into the source of this convergence, it is useful to further segregate the frequencies of anomalous warm months. Figure 2C shows the observed and simulated frequencies of anomalously warm months as a function of latitude for the CMIP3 project. These simulations

overestimate the frequencies of anomalous warm months everywhere but in the Arctic (>60°N), most notably so in the tropics and southern hemisphere. However, the observed frequencies lie within the model simulation envelope except in the midlatitudes of the southern hemisphere. Figure 2D shows a similar result for the CMIP5 project. The hypothesized selection bias favoring simulations that warm more in the Arctic has reduced the simulation spread for all latitudes. As a result, the observed frequencies now lie on the fringes of the model envelope for the bulk of the tropics as well as the southern hemisphere. Curiously, outside of the Arctic, the CMIP5 simulation frequencies of anomalously warm months appear to be converging in the vicinity of the CMIP3/CMIP3 + model ensemble means. Why this should be the case remains obscure, as it is highly unlikely that this metric was used to 'tune' the CMIP5 simulations in any sense.

4. Discussion and Conclusions

[18] It is beyond this (and probably any) study to explain in detail why this apparent convergence towards some common solution is emerging among these CMIP5 project climate change simulations. However, in making this observation of model simulation behavior between two different model generations, it is vital to note the difference between the behavior shown in Figures 1 and 2 and constraining model simulations to fit the observed time evolving climate trajectory. The latter can be argued to be not only valid but ultimately valuable, provided sufficient information is provided regarding how such constraints are implemented [*Knutti*, 2008]. In contrast, the situation here with convergence apparently rooted in the desire to capture one particular regional signature of climate warming is difficult to justify. While the observed Arctic warming is spectacular and important, it is unclear why it is more important from the perspective of the evolution of the overall climate system than the relatively modest warming in the tropics and southern hemisphere. It is unclear whether the CMIP5 simulations are even getting the reason for the actual Arctic warming correct, as they are inconsistent with the strong Arctic warming but only modest warming in the Northern Hemisphere midlatitudes and tropics that best describes the recent evolution of the actual climate system.

[19] What can be done about this situation? First, diversity must be re-injected into climate change simulation behavior [*Huybers*, 2010]. Curiously, in going from the CMIP3 to the CMIP5 projects, not only model simulations that are anomalously weak in their climate warming but also those that are anomalously strong in their warming are suppressed. As a result, the proverbial 'marketplace of ideas' about how climate change has and will continue to occur has shrunk. Instead of 38 unique models, each responding differently to increased anthropogenic forcing, in the CMIP5 project, climate simulation is evolving towards a state where there are 38 'unique' models that all respond the same. Whether through re-examination of the radiative forcings that underlie climate change [*Hansen et al*. 2011], the dynamical variability of the models [*Donner and Large*, 2008], the sensitivity of the models to imposed radiative forcings [*Huybers*, 2010; *Andrews et al*., 2012], or the heat uptake of model oceans [*Meehl et al*., 2011], a healthy dose of diversity must somehow be reintroduced into climate simulation enterprise.

[20] **Acknowledgments.** We gratefully acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modeling (WGCM), for making available the CMIP3 and CMIP5 multimodel datasets. Comments from two anonymous reviewers significantly improved the scope and focus of this work.

References

Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012), Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, *Geophys. Res. Lett.*, *39*, L09712, doi:10.1029/2012GL051607.

Arndt, D. S., M. O. Baringer, and M. R. Johnson (2010), State of the climate in 2009, *Bull. Am. Met. Soc.*, *91*, s1–s222.

Dee, D. P. et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. Roy. Met. Soc.*, *137*, 553–597, doi:10.1002/qj.828.

Donner, L. J., and W. G. Large (2008), Climate modeling, *Ann. Rev. Env. Res.*, *33*, 1–17.

Feynman, R., and R. Leighton (1985), *Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character*, W.W. Norton, New York.

Hansen, J., M. Sato, and R. Ruedy (2012), Perceptions of climate change, *Proc. Nat. Acad. Sci.*, doi:10.1073/pnas.1205276109.

Hansen, J., M. Sato, P. Kharecha, and K. von Schuckmann (2011), Earth's energy imbalance and implications, *Atmos. Chem. Phys.*, *11*, 13421–13449, doi:10.5194/acp-11-13421-2011.

Huybers, P. (2010), Compensation between model feedbacks and curtailment of climate sensitivity, *J. Clim.*, *23*, 3009–3018.

Knutti, R. (2008), Why are climate models reproducing the global surface warming so well?, *Geophys. Res. Lett.*, *35*, L18704, doi:10.1029/2008GL031932.

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

Meehl, G. A., J. M. Arblaster, J. T. Fasullo, A. Hu, and K. E. Trenberth (2011), Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods, *Nat. Clim. Change*, *1*, 360–364, doi:10.1038/nclimate1229.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, *J. Geophys. Res.*, *117*, D08101, doi:10.1029/2011JD017187

Nickerson, R. S. (1998), Confirmation bias; A ubiquitous phenomenon in many guises, *Rev. Gen. Psych.*, *2*(2), 175–220, doi:10.1037/1089-2680.2.2.175.

Poletiek, F. (2001), *Hypothesis-testing behaviour*, Psychology Press, Hove UK.

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

Smith, L. A. (2002), What might we learn from climate forecasts?, *Proc. Nat. Acad. Sci.*, *99*, 2487–2492.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An Overview of CMIP5 and the experiment design, *Bull. Am. Met. Soc.*, doi:10.1175/BAMS-D-11-00094.1.

Trenberth, K. E., and J. Fasullo (2012), Climate extremes and climate change: The Russian heat wave and other climate extremes of 2012, *J. Geophys. Res.*, *117*, D17103, doi:10.1029/2012JD018020.