AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2014GL059205

Key Points:
- Correlation magnitude is not sufficient proof of predictive skill
- Significance testing is complicated by model nonindependence in ensembles
- The best predictors of climate change are related to the Southern Ocean

Supporting Information:
- Readme
- Tables S1 and S2 and Figure S1

Correspondence to:
P. M. Caldwell,
caldwell19@llnl.gov

Citation:
Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson (2014), Statistical significance of climate sensitivity predictors obtained by data mining, *Geophys. Res. Lett.*, *41*, 1803–1808, doi:10.1002/2014GL059205.

Received 3 JAN 2014
Accepted 13 FEB 2014
Accepted article online 17 FEB 2014
Published online 7 MAR 2014

Statistical significance of climate sensitivity predictors obtained by data mining

Peter M. Caldwell[1], Christopher S. Bretherton[2], Mark D. Zelinka[1], Stephen A. Klein[1], Benjamin D. Santer[1], and Benjamin M. Sanderson[3]

[1]Lawrence Livermore National Laboratory, Livermore, California, USA, [2]Department of Atmospheric Sciences, University of Washington, Seattle, Washington, USA, [3]National Center for Atmospheric Research, Boulder, Colorado, USA

Abstract Several recent efforts to estimate Earth's equilibrium climate sensitivity (ECS) focus on identifying quantities in the current climate which are skillful predictors of ECS yet can be constrained by observations. This study automates the search for observable predictors using data from phase 5 of the Coupled Model Intercomparison Project. The primary focus of this paper is assessing statistical significance of the resulting predictive relationships. Failure to account for dependence between models, variables, locations, and seasons is shown to yield misleading results. A new technique for testing the field significance of data-mined correlations which avoids these problems is presented. Using this new approach, all 41,741 relationships we tested were found to be explainable by chance. This leads us to conclude that data mining is best used to identify potential relationships which are then validated or discarded using physically based hypothesis testing.

1. Introduction

Humans have always been fascinated with predicting the future. Making accurate predictions can be extremely difficult, but the payoffs for success can be huge. Predicting changes to Earth's climate over the next hundred years and identifying how humans can influence this future is perhaps the most important prediction problem of our time. But as with most high stakes prediction exercises, understanding climate change is not easy.

The main source of difficulty is that climate responds to complex interactions between weakly understood nonlinear processes. To accommodate this complexity, climate predictions are typically made with global climate models (GCMs) which distill our best understanding of climate processes into numerical models. Unfortunately, independently developed GCMs yield substantially different predictions of future climate [*Flato et al.*, 2013]. This disagreement provides a lower bound on our uncertainty about the magnitude of global warming. Additionally, spread in predictions by successive generations of GCMs does not seem to be decreasing [*Manabe and Wetherald*, 1967; *Andrews et al.*, 2012; *Bony et al.*, 2013].

Because climate change is so important and GCM spread is still substantial, considerable effort has been directed toward finding alternative methods for making climate predictions, particularly for equilibrium climate sensitivity (ECS, the change in global average equilibrium surface air temperature due to doubling CO_2). Empirical estimates of ECS have been obtained from a variety of sources, such as surface temperature changes over the instrumental thermometer record [*Gillett et al.*, 2012], from temperature changes over ice age timescales [*Hoffert and Covey*, 1992], and from temperature responses to large volcanic eruptions [*Wigley et al.*, 2005; *Knutti and Hegerl*, 2008]. The uncertainties in such estimates are large, however, and arise from the combined effects of uncertainties in both the instrumental/proxy data and in the magnitude of key natural and anthropogenic external forcings (such as aerosol and solar insolation) over the last few centuries. In addition, greenhouse gases induce different spatial and temporal feedback response patterns than other forcings [*Hansen et al.*, 1997], which complicate interpretation of ECS estimated from previous climates. See Box 12.2 of *Collins et al.* [2013] for a more comprehensive discussion of techniques for estimating ECS.

An increasingly popular approach is to identify *emergent constraints*—currently observable quantities which serve as good predictors of GCM response to changes in CO_2. This task is made easier by the availability of output from a series of coupled model intercomparison projects (hereafter denoted CMIPn for generation

number n). These projects provide researchers with access to data from dozens of climate models developed by many different groups worldwide. Using such archives, strong correlation to ECS has been identified for the seasonal cycle of near-surface temperature [*Covey et al.*, 2000; *Knutti et al.*, 2006], the gradient between tropical and midlatitude mean cloud amount [*Volodin*, 2008], Southern Hemisphere radiation [*Trenberth and Fasullo*, 2010], subtropical midtropospheric relative humidity [*Fasullo and Trenberth*, 2012], the lapse rate of tropical humidity [*Sherwood et al.*, 2014], and various other radiative quantities [*Huber et al.*, 2011].

A natural extension to the above-mentioned studies is to automate the process of identifying current climate quantities with skill at predicting ECS. In this study, we correlate climate sensitivity from 28 CMIP5 models (calculated as described in the supporting information) against 41,741 vectors of current climate data from CMIP5. As described in the supporting information, these vectors (hereafter known as *fields*) sample the mean, interannual standard deviation, and seasonal amplitude of 48 variables over a range of latitude bands, seasons, and vertical levels. This work expands upon *Huber et al.* [2011], which searched through 1700 permutations of climate metrics and geographical regions in 28 CMIP3 radiation variables for correlations with ECS. However, while the previous effort struggled to establish the statistical significance of their results, our study focuses exclusively on identifying which (if any) correlations are significant.

Establishing the significance of data-mined relations is difficult because sampling unusual events becomes likely as sample size increases. For example, if 100 independent tests are conducted at 5% significance level, five of those tests are expected to appear significant by chance alone. In meteorology and climate applications, significance tests for ensembles of statistical relationships are known as *field significance* tests [*Livezey and Chen*, 1983]. Several methods have been proposed for establishing the significance of relationships in portions of spatially and temporally varying arrays [*von Storch*, 1982; *Livezey and Chen*, 1983; *Wilks*, 2006; *DelSole and Yang*, 2011]. Searching the CMIP archive for correlations is different than analyzing the spatial maps of previous studies because our fields are correlated not just across space and time but also across variables and models. This makes computing the effective sample size or constructing analogous bootstrapped data sets difficult. Additionally, the statistics in our ensemble are constructed from nonindependent models, a feature shown in section 3 to play a central role in determining field significance. While previous studies have noted the nonindependence of CMIP models [*Knutti*, 2010; *Pennell and Reichler*, 2011; *Knutti et al.*, 2013], ours is the first we know of which accounts for all of these complex dependencies in the determination of field significance.

Section 2 provides a simple but flawed approach to analyzing data-mined correlations. It serves to introduce the basic methodology used throughout the paper and to set the stage for section 3, which identifies the problems with this initial approach. A more *suitable* methodology is presented in section 4. Section 5 summarizes the results and places them in a broader context.

2. A Naïve Approach to Field Significance Testing

Figure 1 shows the histogram of the magnitudes of correlations between intermodel differences in ECS and each of the 41,741 current climate fields in our data set. Here and elsewhere, quantities have been normalized by subtracting their intermodel mean value and dividing the result by their intermodel standard deviation. Blue bars depict correlations with ECS and red bars show correlations with independent, uniformly distributed random data. We perform our null hypothesis calculation by replacing ECS rather than the fields with random data in order to preserve the rich structure of correlations between fields and across models. Replicating this structure is shown below to be essential to the formation of an appropriate significance test.

In both the ECS and random data cases, weak correlations are prevalent and strong correlations are relatively rare. This behavior is unsurprising but has a profound impact on the type of relationship which can be identified via data mining. Weak but physically meaningful relationships are hidden under a mountain of chance correlations, while strong correlations stand out because they are relatively rare. This leads us to focus on the strongest correlations. However, deciding that a large correlation is real is equivalent to concluding that a single aspect of current climate controls a dominant fraction of global warming. We consider this unlikely given the incredible complexity of the climate system. Inability to identify weak but meaningful correlations is a drawback to data mining.

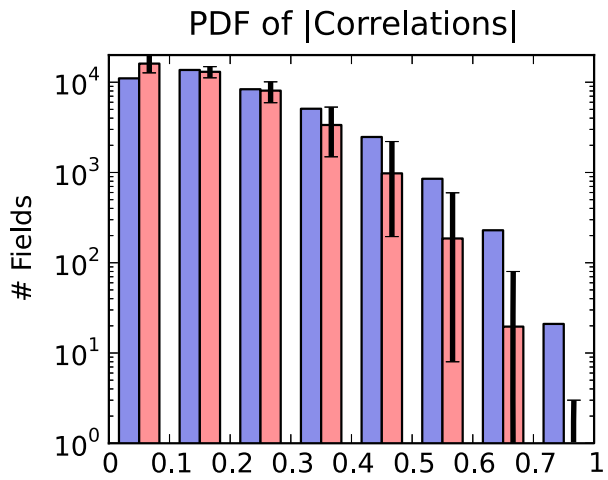

Figure 1. Histogram of $\{|\text{corr}(X, Y)|\}$ for $X \in$ CMIP5 fields and $Y =$ ECS (blue bars) or random data (red bars)}. For random data, error bounds are 5 and 95 percentile values computed by drawing a new random vector and recomputing the histogram 1000 times.

The key feature of Figure 1 is that the number of correlations greater than 0.7 in absolute value (hereafter known as the set of *strongly correlated* fields) in the CMIP5 archive exceeds the 95th percentile value generated with randomized ECS vectors. This suggests that several of the largest correlations identified are not just statistical artifacts but instead reflect real, physically based relationships between CMIP5 fields and ECS. This result, however, is based on several incorrect assumptions which are described below.

3. Problems With the Naïve Approach

One issue with the above approach is that cross-field independence was not considered in selecting which fields to correlate with ECS. As a result, the number of fields strongly correlated with ECS is highly dependent on arbitrary choices about how many fields in the archive express a particular pattern of cross-model variability. For example, if intermodel variations in ECS and tropical free-tropospheric temperature (which is strongly related to many other tropical fields) were highly correlated, the number of strong correlations would be very large. If ECS was instead strongly correlated with surface sensible heat flux near the North Pole (a quantity less well correlated with other fields), fewer strong correlations with ECS would be identified. The impact of removing closely related fields is tested in Figure 2a by identifying the best correlated pair of fields and randomly removing one of them, then iterating until the desired number of remaining fields is left. As fields are removed there are fewer opportunities for large correlations, and as a result, the number of strong correlations decreases monotonically. As in Figure 1, the number of strong correlations is significant at the 5% level wherever the value for the CMIP results is not encompassed by the error bars. Decreasing redundancy between fields is seen to affect the number of strongly correlated fields but does not (in this case) impact significance.

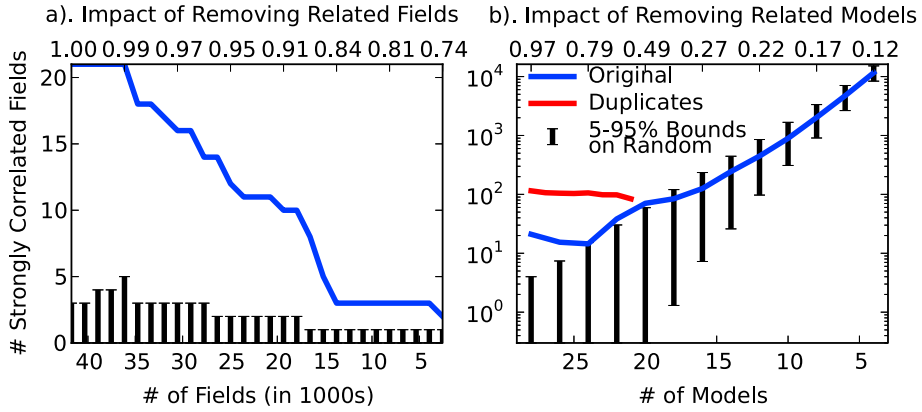

Figure 2. Number of fields in the CMIP5 archive strongly correlated with ECS (defined as having $|\text{corr}(X, \text{ECS})| > 0.7$) as a function of (a) number of fields retained and (b) number of models retained. Error bars are 5–95% bounds on the number of CMIP5 fields strongly correlated with a random data vector. The red line in Figure 2b shows the impact of reducing model count to 20, then increasing model count by randomly adding copies of retained models. Numbering above plots gives the maximum correlation between fields or models for given field or model count. See text for details.

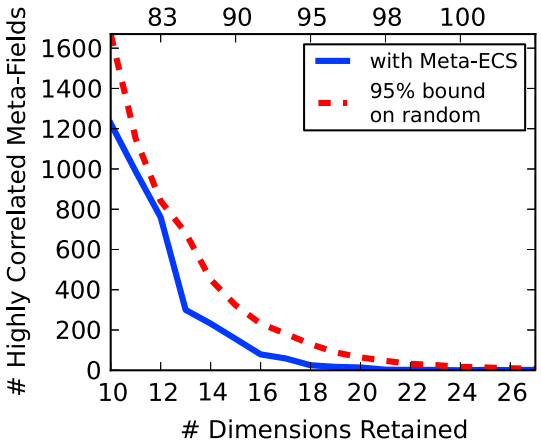

Figure 3. Number of meta-fields correlated with meta-ECS at >0.7 in absolute value as a function of the number of dimensions retained. The red line is the 95th percentile number of correlations expected by chance based on 1000 trials. Numbering at the top of the figure gives the percent of the variance in the intermodel correlation matrix explained by the retained EOFs.

A more fundamental problem with the original analysis is that it assumes CMIP5 models are independent. Figure 2b illustrates the impact of removing the most strongly related models (following the same methodology as used for Figure 2a). To minimize the impact of randomly choosing which model of the strongest-correlated pair is removed, model reduction is performed 20 times and the results displayed are the average over this sampling. In general, using more models results in fewer strong correlations. This reflects the simple fact that the probability of matching a pattern by chance decreases as the complexity of the pattern increases. Artificially inflating the model count by including some models multiple times, however, does *not* change the number of strong correlations. This is illustrated by removing the eight strongest-correlated models, then reinflating model count by randomly adding copies of the retained models. The red line in Figure 2b shows averages over 20 repetitions of this process. Lack of sensitivity to redundant models is unsurprising: it is trivial to show that including all models multiple times has no effect on the histogram of correlations. This insensitivity of the histogram of correlations to inclusion of redundant models explains why the number of strong correlations in the CMIP5 data (blue line in Figure 2b) is practically unchanged when model count is reduced from 28 to 22 by removing the strongest-related models. The number of strong correlations in our null distribution from section 2 does, however, continue to decrease as the number of closely related models increases. This is because randomization replaces copied ECS values with independent data. In short, our null hypothesis implicitly assumed models were independent, so it failed to reproduce the artificial increase in strong correlations with ECS caused by using related models.

4. A Better Approach to Field Significance Testing

A crude significance test correcting for related models can be made by noting that the number of strong correlations in Figure 2b remains roughly constant until $\geq$4 models are removed, so there must be $\sim$24 independent models (see supporting information for a more comprehensive discussion of this point). Using 24 models, the number of strong correlations from CMIP data is roughly equal to the 95th percentile for random data.

A more suitable method of field significance testing is summarized below (and described in more detail in the supporting information). The basic idea is to replace the original, related models with linearly independent "meta-models" which are formed from linear combinations of the original models using the empirical orthogonal functions (EOFs) of the matrix of correlations between models. We then reduce model dimensionality to account for there being fewer than 28 independent models. Finally, ECS and field vectors for the meta-models (hereafter meta-ECS and meta-fields) are constructed as linear combinations of the original data. Analyzing correlations between meta-fields and meta-ECS is equivalent to performing the original significance test using independent models and fields. Figure 3 shows the number of meta-fields correlated strongly with meta-ECS as a function of the assumed number of independent models in the data set. Results are shown for all reasonable independent model counts to avoid making somewhat arbitrary decisions about which models are independent. The null hypothesis (that the number of fields strongly correlated with ECS is commensurate with the number expected by chance) is tested by replacing meta-ECS with independent, uniformly distributed random numbers. The number of strong correlations using the CMIP5 data (blue line) is always less than the 95th percentile from 1000 tests using random data (red line), so we conclude that the number of strong correlations with ECS is not significant at the 5% level regardless of how many independent models are assumed.

5. Discussion and Conclusions

Because of the scientific and societal importance of reducing uncertainties in climate sensitivity estimates, the widespread availability of simulation output from many dozens of climate models, and emerging techniques for quickly sorting through massive data sets, archives such as CMIP5 are prime candidates for data mining. Our study serves as a first foray into this realm. The major conclusion of this work is that credible statistical significance testing of mined data requires careful consideration of relationships between samples. Our analysis shows that none of the 41,741 fields tested have skill at predicting ECS beyond what is reasonably expected by chance.

In this context it is worth revisiting earlier attempts to link ECS with currently observable quantities. Our work demonstrates that extremely large correlations are expected by chance in the CMIP archives, so simply identifying a strong correlation is not sufficient proof that a real physical relationship exists. Instead, emergent constraints require a convincing physical explanation to be credible. If that physical explanation is what prompted the significance test, a simple *t* test is sufficient. If the relationship precedes the explanation (as in data mining), the probability of getting unusual events by chance becomes large and the ability of statistics to separate spurious and meaningful relationships drops. Previously identified emergent constraints which lack a clear physical explanation should be treated with caution. A necessary but not sufficient condition for their continued acceptance is that the suggested relationship should continue to hold in other ensembles of models [*Masson and Knutti*, 2013; *Klocke et al.*, 2011]. This criterion is not met in *Fasullo and Trenberth* [2012], where subtropical and monsoon region relative humidity are found to be good predictors of ECS in CMIP3 but not in CMIP5 (see their Figure S4).

In light of the importance of a physical explanation, it may be easier to establish emergent constraints for individual feedbacks than for ECS since the former are more closely connected to physical processes. Several constraints of this type have already been proposed [*Hall and Qu*, 2006; *Qu et al.*, 2013; *Cox et al.*, 2013].

Although data mining was not found to be useful for identifying skillful predictors in CMIP5, we do not mean to imply that no such relationships exist or that data mining has no place in climate science. Even though statistical significance of correlations could not be established here using data mining, data mining does give us a comprehensive list of strong correlations (see Table S2). This serves as a great starting point for identifying real predictive relationships.

Acknowledgments

We would like to thank two anonymous reviewers for very useful comments. We would also like to acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the World Climate Research Program's Working Group on Coupled Modeling for their roles in making available the CMIP5 multimodel data set. Support of the CMIP5 data set is provided by the U.S. Department of Energy (DOE) Office of Science. This work was performed under the auspices of DOE by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. C.S. Bretherton is funded by NASA grant NNX09AH73G. B.M. Sanderson is supported by the Office of Science, Biological and Environmental Research, U.S. Department of Energy, cooperative agreement DE-FC02-97ER62402, and the National Center for Atmospheric Research which is sponsored by the National Science Foundation. All other authors are supported by DOE's Regional and Global Climate Modeling Program. Data used in this study are publicly available at http://cmip-pcmdi.llnl.gov.

The Editor thanks two anonymous reviewers for their assistance in evaluating this paper.

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.

Bony, S., G. Bellon, D. Klocke, S. Sherwood, S. Fermepin, and S. Denvil (2013), Robust direct effect of carbon dioxide on tropical circulation and regional precipitation, *Nature*, *6*, 447–451, doi:10.1038/ngeo1799.

Collins, M., et al. (2013), Long-term climate change: Projections, commitments and irreversibility, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 1029–1136, Cambridge Univ. Press, Cambridge, U. K., and New York.

Covey, C., et al. (2000), The seasonal cycle in coupled ocean-atmosphere general circulation models, *Clim. Dyn.*, *16*, 775–787.

Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, and C. M. Luke (2013), Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, *Nature*, *494*, 341–344, doi:10.1038/nature11882.

DelSole, T., and X. Yang (2011), Field significance and regression patterns, *J. Clim.*, *24*, 5094–5107.

Fasullo, J., and K. Trenberth (2012), A less cloudy future: The role of subtropical subsidence in climate sensitivity, *Science*, *338*, 792–794, doi:10.1126/science.1227465.

Flato, G., et al. (2013), Evaluation of climate models, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 741–866, Cambridge Univ. Press, Cambridge, U. K., and New York.

Gillett, N. P., V. K. Arora, G. M. Flato, J. F. Scinocca, and K. von Salzen (2012), Improved constraints on 21st-century warming derived using 160 years of temperature observations, *Geophys. Res. Lett.*, *39*(1), L01704, doi:10.1029/2011GL050226.

Hall, A., and X. Qu (2006), Using the current seasonal cycle to constrain snow albedo feedback in future climate change, *Geophys. Res. Lett.*, *33*, L03502, doi:10.1029/2005GL025127.

Hansen, J., M. Sato, and R. Ruedy (1997), Radiative forcing and climate response, *J. Geophys. Res.*, *102*(D6), 6831–6864.

Hoffert, M. I., and C. Covey (1992), Deriving global climate sensitivity from palaeoclimate reconstructions, *Nature*, *360*, 573–576.

Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti (2011), Constraints on climate sensitivity from radiation patterns in climate models, *J. Clim.*, *24*, 1034–1052, doi:10.1175/2010JCLI3403.1.

Klocke, D., R. Pincus, and J. Quaas (2011), On constraining estimates of climate sensitivity with present-day observations through model weighting, *J. Clim.*, *24*, 6092–6099.

Knutti, R. (2010), The end of model democracy?, *Clim. Change*, *102*, 395–404.

Knutti, R., and G. C. Hegerl (2008), The equilibrium sensitivity of the Earth's temperature to radiation changes, *Nat. Geosci.*, *1*, 735–743.

Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth (2006), Constraining climate sensitivity from the seasonal cycle in surface temperature, *J. Clim.*, *19*, 4224–4233.

Knutti, R., D. Masson, and A. Gettelman (2013), Climate model genealogy: Generation CMIP5 and how we got there, *Geophy. Res. Let.*, *40*, 1194–1199, doi:10.1002/grl.50256.

Livezey, R. E., and W. Chen (1983), Statistical field significance and its determination by Monte Carlo techniques, *Mon. Weather Rev.*, *111*, 45–59.

Manabe, S., and R. T. Wetherald (1967), Thermal equilibrium of the atmosphere with a given distribution of relative humidity, *J. Atmos. Sci.*, *50*, 241–259.

Masson, D., and R. Knutti (2013), Predictor screening, calibration, and observational constraints in climate model ensembles: An illustration using climate sensitivity, *J. Clim.*, *26*, 887–898.

Pennell, C., and T. Reichler (2011), On the effective number of climate models, *J. Clim.*, *24*, 2358–2367.

Qu, X., A. Hall, S. A. Klein, and P. M. Caldwell (2013), On the spread of changes in marine low cloud cover in climate model simulations of the 21st century, *Clim. Dyn.*, doi:10.1007/s00382-013-1945-z.

Sherwood, S. C., S. Bony, and J.-L. Dufresne (2014), Spread in model climate sensitivity traced to atmospheric convective mixing, *Nature*, *505*(7481), 37–42, doi:10.1038/nature12829.

Trenberth, K. E., and J. T. Fasullo (2010), Simulation of present day and 21st century energy budgets of the southern oceans, *J. Clim.*, *23*, 440–454.

Volodin, E. M. (2008), Relation between temperature sensitivity to doubled carbon dioxide and the distribution of clouds in current climate models, *Izv. Atmos. Oceanic Phys.*, *44*, 288–299.

von Storch, H. (1982), A remark on Chervin-Schneider's algorithm to test significance of climate experiments with GCMs, *J. Atmos. Sci.*, *39*, 187–189.

Wigley, T. M. L., C. M. Ammann, B. D. Santer, and S. C. B. Raper (2005), The effect of climate sensitivity on the response to volcanic forcing, *J. Geophys. Res.*, *110*, D09107, doi:10.1029/2004/JD005557.

Wilks, D. S. (2006), On "field significance" and the false discovery rate, *J. Appl. Meteor. Climatol.*, *45*, 1181–1189.