# Benchmark forecasts for climate change

Green, Kesten C and Armstrong, J Scott and Soon, Willie

12 December 2008

# Benchmark Forecasts for Climate Change

**Kesten C. Green**
Business and Economic Forecasting, Monash University, Vic 3800, Australia.
Contact: PO Box 10800, Wellington 6143, New Zealand.
kesten@kestencgreen.com; T +64 4 976 3245; F +64 4 976 3250

**J. Scott Armstrong**
The Wharton School, University of Pennsylvania
747 Huntsman, Philadelphia, PA 19104
armstrong@wharton.upenn.edu; jscottarmstrong.com; T +1 610 622 6480

**Willie Soon**
Harvard-Smithsonian Center for Astrophysics, Cambridge MA 02138
wsoon@cfa.harvard.edu; T +1 617 495 7488

December 12, 2008

## ABSTRACT

We assessed three important criteria of forecastability—simplicity, certainty, and variability. Climate is complex due to many causal variables and their variable interactions. There is uncertainty about causes, effects, and data. Using evidence-based (scientific) forecasting principles, we determined that a naïve "no change" extrapolation method was the appropriate benchmark. To be useful to policy makers, a proposed forecasting method would have to provide forecasts that were substantially more accurate than the benchmark. We calculated benchmark forecasts against the UK Met Office Hadley Centre's annual average thermometer data from 1850 through 2007. For 20- and 50-year horizons the mean absolute errors were 0.18°C and 0.24°C. The accuracy of forecasts from our naïve model is such that even perfect forecasts would be unlikely to help policy makers. We nevertheless evaluated the Intergovernmental Panel on Climate Change's 1992 forecast of 0.03°C-per-year temperature increases. The small sample of errors from ex ante forecasts for 1992 through 2008 was practically indistinguishable from the naïve benchmark errors. To get a larger sample and evidence on longer horizons we backcast successively from 1974 to 1850. Averaged over all horizons, IPCC errors were more than seven-times greater than errors from the benchmark. Relative errors were larger for longer backcast horizons.

Key words: backcasting, climate model, decision making, ex ante forecasts, out-of-sample errors, predictability, public policy, relative absolute errors, unconditional forecasts.

## Introduction

One of the principles of scientific forecasting is to ensure that a series can be predicted (Armstrong, 2001, Principle #1.4). We applied the principle to long-term forecasting of global mean temperatures by examining the unconditional ex ante forecast errors from a naïve benchmark model. By ex ante forecasts, we mean forecasts for periods that were not taken into account when the forecasting model was developed—it is trivial to construct a model that fits known data better than a naïve model can.

Benchmark errors are the standard by which to determine whether alternative scientifically-based forecasting methods can provide useful forecasts. When benchmark errors are large, it is possible that alternative methods would provide useful forecasts. When benchmark errors are small, it is less likely that other methods will be able to provide improvements in accuracy that are useful to decision makers.

## Conditions of forecastability

By forecastability we mean the ability to improve upon a naïve benchmark model. Three important conditions of forecastability are variability, simplicity, and certainty.

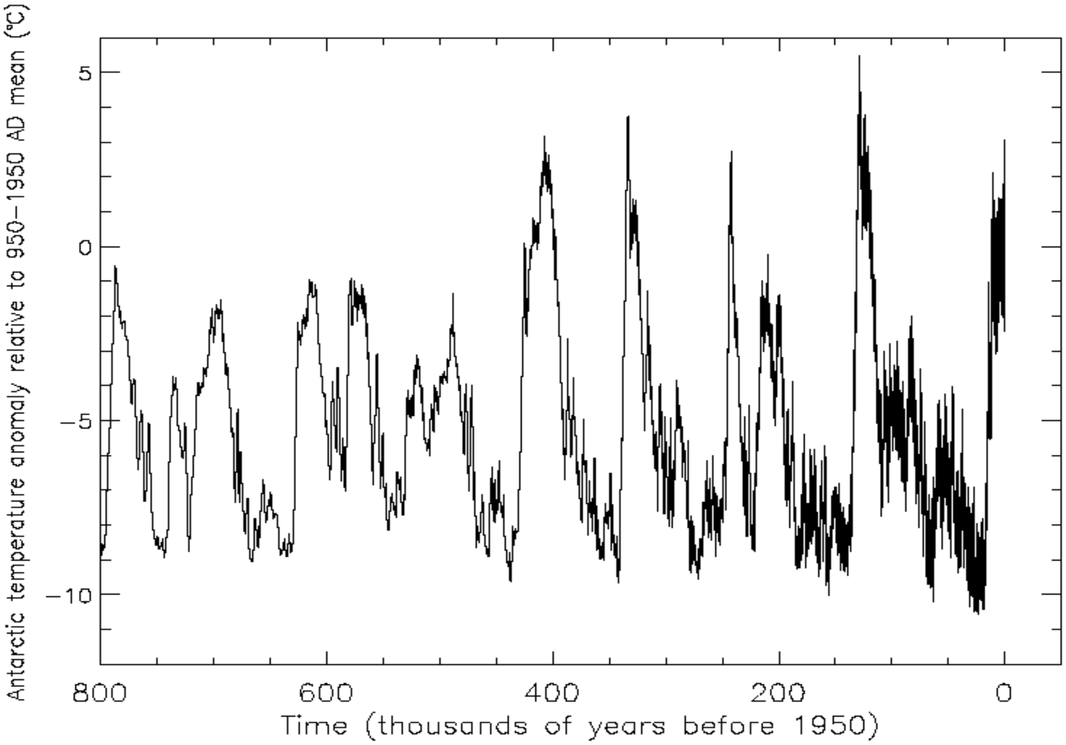
### *Variability*

The first step in testing whether a forecasting method can help is to check for variability. If little or no variability is expected, there is no need to make a forecast

In the case of global mean temperatures, warnings since 1990 from the Intergovernmental Panel on Climate Change (IPCC) and others (Hansen 2008) that we are experiencing dangerous manmade global warming suggest variability. Indeed, when we examined local, regional and global mean temperature data we found that changes are common. For example, Exhibit 1 displays Antarctic temperature data from the ice-core record for the 800,000 years to 1950. The data are in the form of temperature, relative to the average for the last one-thousand-years of the record (950 to 1950 AD), in degrees Celsius. The data show long-term variations. The three most recent values are roughly 1 to 3°C warmer than the reference thousand-year average, which is at 0°C in the graph. Moreover, there was high variability around trends and the trends were unstable over all time periods. In other words, trends appear to be positive about as often as they were negative.

INSERT EXHIBIT 1 ABOUT HERE

**800,000-year Record of Temperature Change**



*Simplicity*

To the extent that a situation is complex, it is more difficult to forecast. This is especially important when complexity is high relative to the variability in the series. For example, daily movements in stock market prices involve complex interactions among many variables. As a consequence daily stock price movements are characterized as a random walk. The naive no-change benchmark method for forecasting stock prices has defeated alternative investment strategies. Attempts to improve upon this model have led to massive losses on occasion, such as with the failure of hedge fund Long-Term Capital Management in the late-1990s.

Climate change is also subject to many interacting variables. The Sun is clearly one important influence on Earthly temperatures. The Sun's intensity varies, the Earth-Sun distance varies, and so does the geometrical orientation of the Earth toward the Sun. The approximately 11-year solar activity cycle, for example, is typically associated with a global average temperature range of approximately 0.4°C between the warmest and coldest parts of the cycle, and a much larger range near the poles (Camp and Tung 2007). Variations in the irradiance of the Sun over decades and centuries also influence the Earth's climate (Soon 2009). Other influences on both shorter and longer-term temperatures include the type and extent of clouds, the extent and reflectivity of snow and ice, ocean currents and the release and absorption of heat by the oceans.

*Certainty*

There is high uncertainty with respect to the direction and magnitude of the various postulated causal factors.

Those who warn of dangerous manmade global warming assert that it is being caused by increasing concentrations of carbon dioxide ($CO_2$) in the atmosphere as a result of human emissions. However, the relationship between human emissions and total atmospheric concentrations is not well-understood due to the complexity of global carbon cycling via diverse physical, chemical, and biological interactions among the $CO_2$ reservoirs of the Earth system. For example, 650,000 years of ice core data suggest that atmospheric concentrations of $CO_2$ have *followed* temperature changes by several hundreds to several thousands of years (Soon 2007). Moreover, there are debates among scientists as to whether additions to atmospheric $CO_2$ play a role of any importance in climate change (e.g., Carter et al. 2006; Soon 2007; Lindzen 2009).

There is also uncertainty about temperature series that have been used by the IPCC. These have been challenged on the basis that they are not true global averages, and that they suffer from "heat island" effects whereby weather stations that were once beyond the edge of town have become progressively surrounded by urban development. Other influences on temperature readings include the substitution of electronic thermometers, which are sensitive to heat eddies; the reduction of the number of temperature stations (especially in remote areas); and maintenance associated with the housing of the temperature gauges (the boxes are supposed to be white). Anthony Watts and colleagues have documented problems with weather station readings at surfacestations.org. Analysis by McKitrick and Michaels (2007) suggested that the size of the surface warming in the last two decades of the 20th century was overestimated by a factor of two.

Finally, long time-series of reliable global and regional temperature data and of the host of plausible causal variables are not available.

In sum, two of three important conditions of forecastability are not met: uncertainty and complexity suggest that climate change will have low predictability.

## An appropriate benchmark model

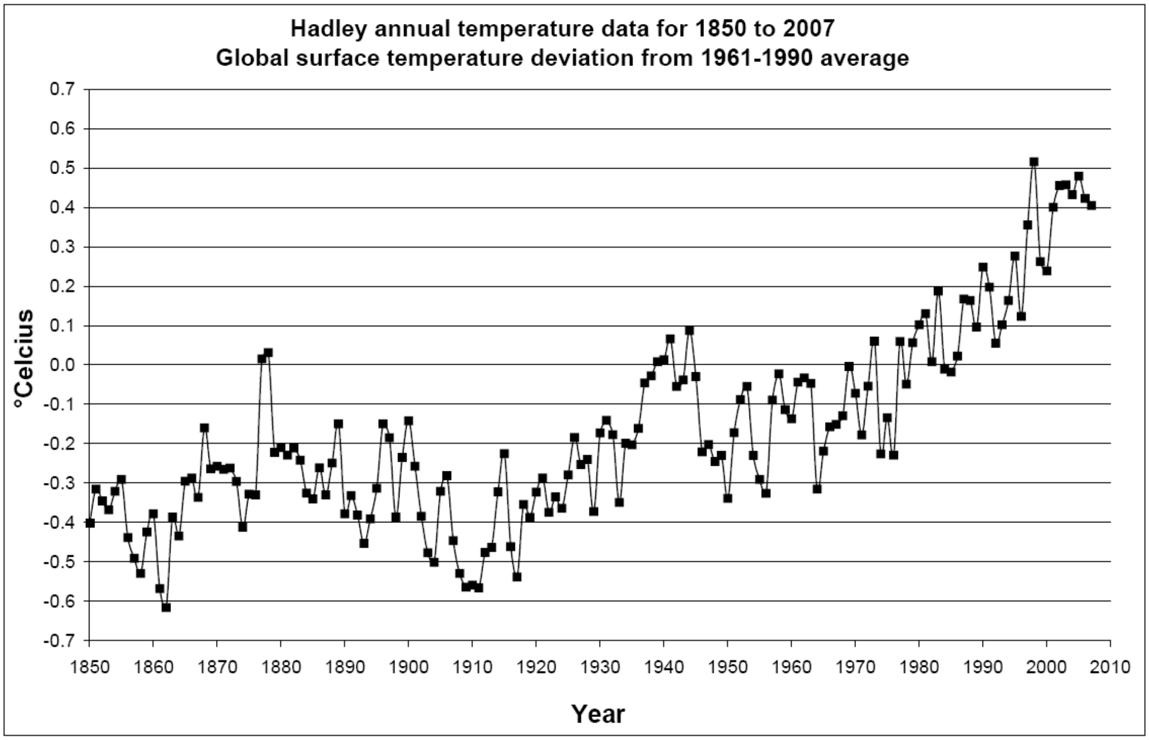
We followed the guidance provided by comparative empirical studies from all areas of forecasting. The guidelines are summarized in Armstrong (2001) and are available on the public service website ForPrin.com.

Given the uncertainty and the complexity of our long-term global average temperature forecasting problem, the lack of agreement among climate scientists on the net directional effects of causal forces, and the lack of consistent long-term trends in the data, the appropriate benchmark is a naïve, no-change, forecasting model.

We used the HadCRUt3 "best estimate" annual average temperature differences from 1850 to 2007 from the U.K. Met Office Hadley Centre (Hadley) [1] to examine the benchmark errors for climate change (Exhibit 2).

[1] Obtained from http://hadobs.metoffice.com/hadcrut3/diagnostics/global/nh+sh/annual on 9 October, 2008.

INSERT EXHIBIT 2



*Errors from the benchmark model*

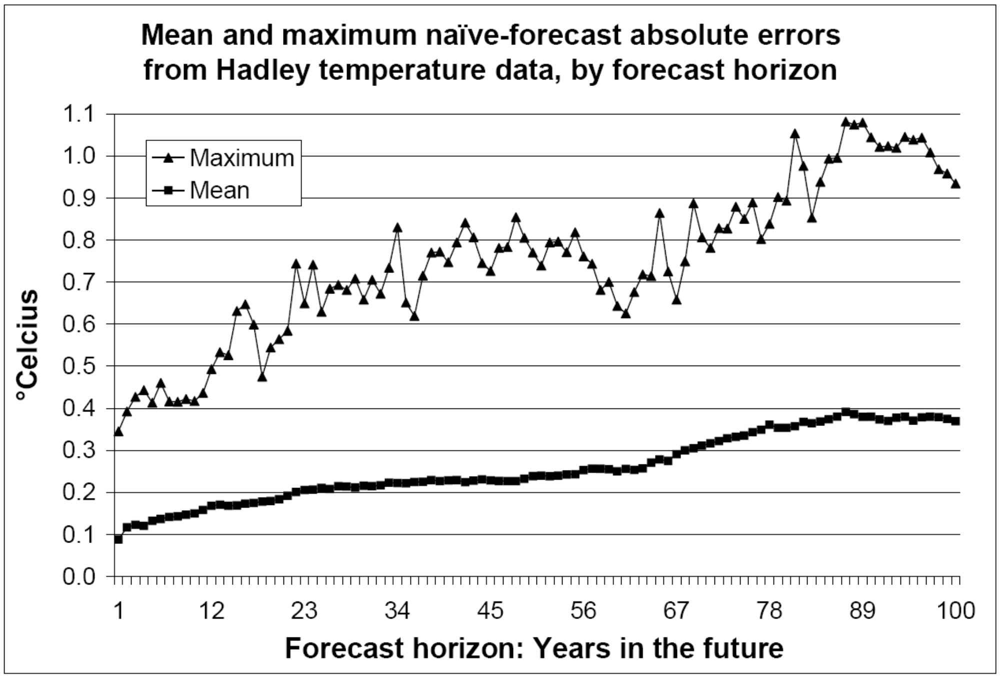
We used each year's mean global temperature as a naïve forecast of each subsequent year and calculated the errors relative to the measurements for those years. For example, the year 1850 temperature measurement from Hadley was our forecast of the average temperature for each year from 1851 through 1950. We calculated the differences between our naïve forecast and the Hadley measurement for each year of this 100-year forecast horizon.

In this way we obtained from the Hadley data 157 error estimates for one-year-ahead forecasts, 156 for two-year-ahead forecasts, and so on up to 58 error estimates for 100-year-ahead forecasts; a total of 10,750 forecasts across all horizons

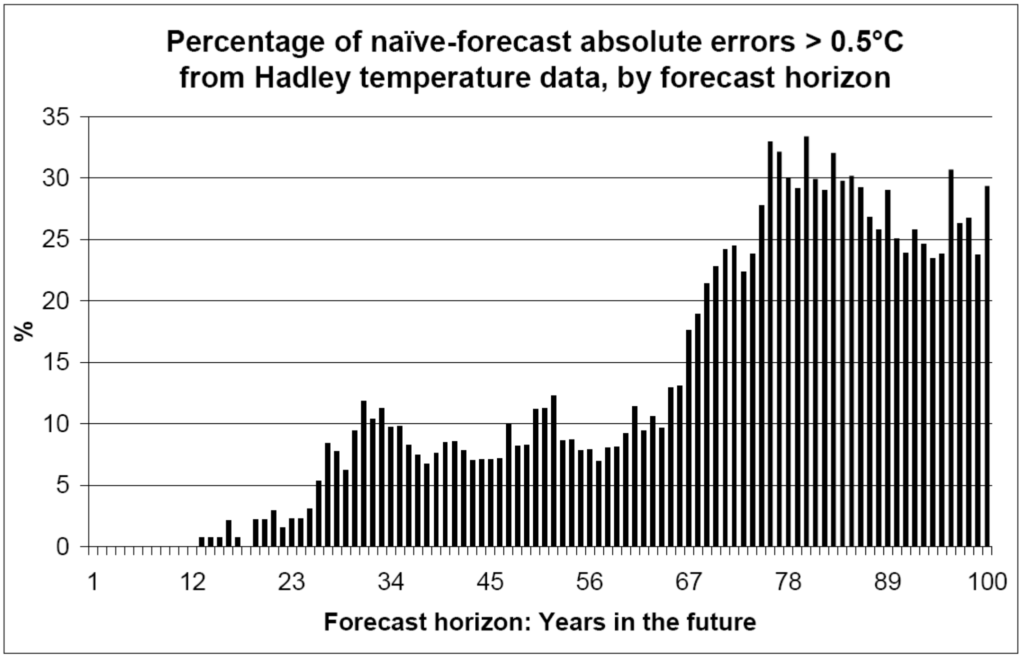
Exhibit 3 shows that mean absolute errors from our naïve model increased from less than 0.1°C for one-year-ahead forecasts to less than 0.4°C for 100-year-ahead forecasts. Maximum absolute errors increased from less than 0.4°C for one-year-ahead forecasts to less than 1.0°C for 100-year-ahead forecasts.

Overwhelmingly, errors were no-more-than 0.5°C, as is shown in Exhibit 4. For horizons less-than-65-years, fewer than one-in-eight of our ex ante forecasts were more than 0.5°C different from the Hadley measurement. All forecasts for horizons up-to-80-years and more than 95% of forecasts for horizons from-81-to-100-years were within 1°C of the Hadley figure. The overall maximum error from all 10,750 forecasts for all horizons was 1.08°C; which was from an 87-year-ahead forecast for the year 1998—the hottest year of a major El Niño cycle.

INSERT EXHIBIT 3



INSERT EXHIBIT 4



## Performance of Intergovernmental Panel on Climate Change projections

As the naïve benchmark model performs so well it is hard to argue what additional benefits public policy makers would get from a better model. Governments did however, via the United Nations, establish the IPCC to search for a better model. The IPCC forecasts provide an opportunity to illustrate the use of our naïve benchmark.

Green and Armstrong (2008) analyzed the IPCC procedures and concluded that they violated 72 of the principles for proper scientific forecasting. For important forecasts, it is critical that all proper procedures are followed. An invalid forecasting method might provide an accurate forecast by chance, but this would not qualify it as an appropriate method. Nevertheless, because the IPCC forecasts influenced major policy decisions, we compare its predictions with our naïve benchmark.

To test any forecasting method, it is necessary to exclude data that were used to develop the model; that is, the testing must be done using out-of-sample data. The most obvious out-of-sample data are the observations that occurred after the forecast was made. There have, however, been only 17 observations of annual global average temperature since the IPCC's 1992 forecasts (including an estimate for 2008) and so we decided to also employ "backcasting".

Dangerous manmade global warming became an issue of public concern after NASA scientist James Hansen testified on the subject to the U.S. Congress on 23 June 1988 (McKibben 2007). The IPCC (2007) authors explain however that "Global atmospheric concentrations of carbon dioxide, methane and nitrous oxide have increased markedly as a result of human activities since 1750" (p. 2). As a consequence we used the Hadley data from 1974 through to the beginning of the series in 1850 for our backcast test.

We used the IPCC's 1992 forecast, which was an update of their 1990 forecast, for our demonstration. The 1992 forecast was for an increase of 0.03°C per year (IPCC 1990 p. xi, IPCC 1992 p.17). We used this forecast because it has had a big influence on policymakers, coming out as it did in time for the Rio Earth Summit, which produced inter alia Agenda 21 and the United Nations Framework Convention on Climate Change. According to the United Nations web page on the Summit [2], "The Earth Summit influenced all subsequent UN conferences…". Using the 1992 forecast also allowed for the longest *ex ante* forecast test. Spreadsheets of our analysis are available at publicpolicyforecasting.com.

There remains the unresolved problem that the IPCC authors knew in retrospect that there had been a broadly upward trend in the Hadley temperature series. From 1850 to 1974 there were 66 years in which the temperature increased from the previous year and 59 in which it declined. There will, therefore, be some positive trend that would provide a better model for the backcast test period than would our naïve benchmark, and so the benchmark is disadvantaged for the period under consideration. In other words, although we treat this as an out-of-sample period, it presumably influenced the thinking of the IPCC experts such that their forecasting model likely fits the 1850 to 1975 trend more closely than it would had they been unaware of the data. Recall, however, that the temperature variations shown by the longer temperature series in Exhibit 1 suggest that there is no assurance that the trend will continue in the future.

[2] http://www.un.org/geninfo/bp/enviro.html

### *Evaluation method*

We followed the procedure that we had used for our benchmark model and calculated absolute errors as the unsigned difference between the IPCC forecast, or backcast, and the Hadley figure for the same year. We then compared these IPCC forecast errors with those from the benchmark model using the cumulative relative absolute error or CumRAE (Armstrong 2001).

The CumRAE is the sum across all forecast horizons of the errors (ignoring signs) from the method being evaluated divided by the equivalent sum of benchmark errors. For example, a CumRAE of 1.0 would indicate that the evaluated-method errors and benchmark errors came to the same total while a figure of 0.8 would indicate indicates that the evaluated-method errors were in total 20% lower than the benchmark's.

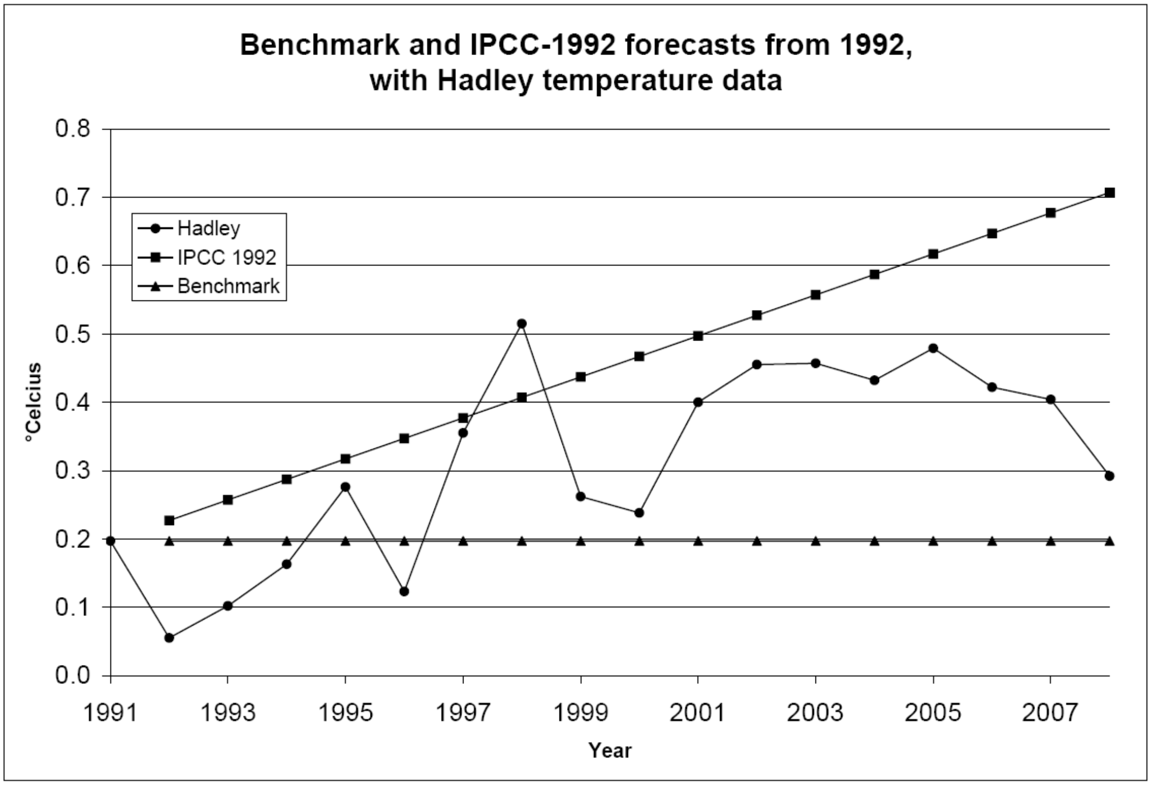
We are concerned about forecasting accuracy by forecast horizon and so calculated error scores for each horizon, and then averaged across the horizons. Thus, the CumRAEs we report are the sum of the mean absolute errors across horizons divided by the equivalent sum of benchmark errors.

### *Forecasts from 1992 through 2008 using 1992 IPCC model*

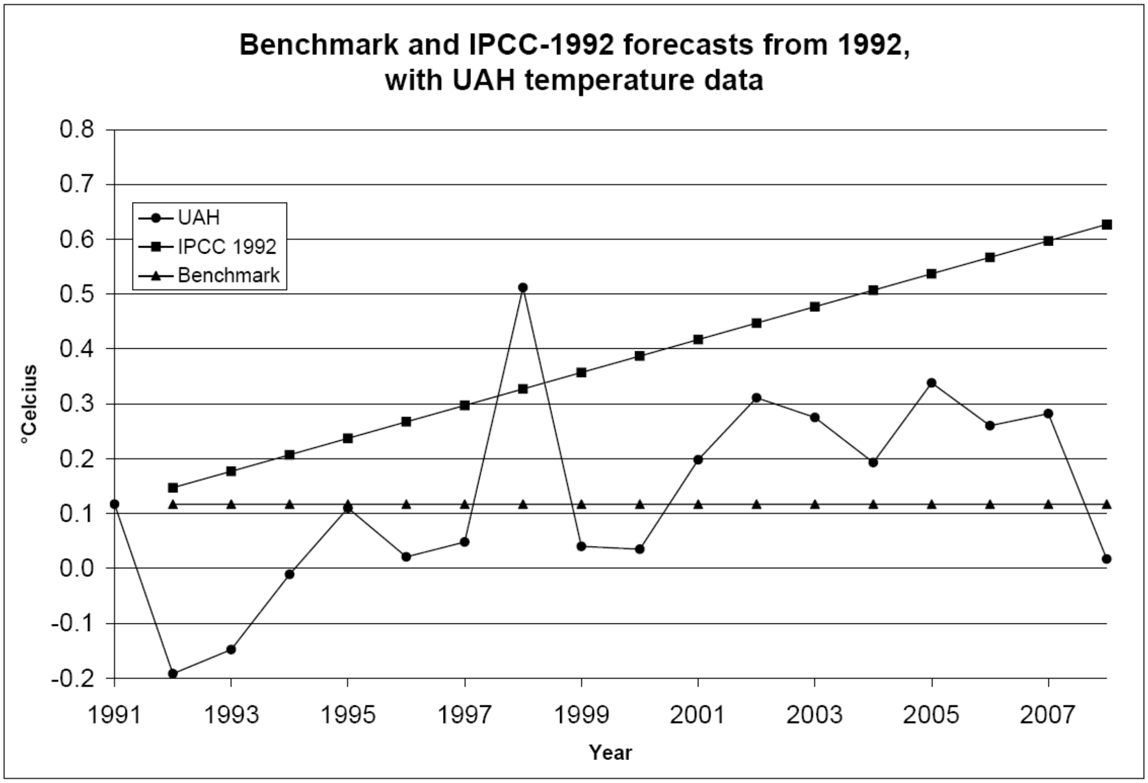
We created an IPCC forecast series from 1992 to 2008 by starting with the 1991 Hadley figure and adding 0.03°C per year. In the case of forecasts, as opposed to backcasts, it is possible to also test the IPCC model against the University of Alabama's data of global near surface temperature measured from satellites using microwave sounding units (UAH), which are available from 1979. We created another forecast series by starting with the 1991 UAH figure.

Benchmarks for the two series were the 1991 Hadley figure and the 1991 UAH figure, respectively, for all years. This process, by including estimates for 2008 from both sources, gave us two small samples of 17 years of out-of-sample forecasts. We found the 1992 IPCC model forecasts were less accurate than the forecasts from our naïve benchmark. When tested against Hadley measures (data plotted in Exhibit 5), IPCC errors were essentially the same as those from our benchmark forecasts (CumRAE 0.98); they were nearly twice as large (CumRAE 1.82) when tested against the UAH satellite measures (Exhibit 6).

INSERT EXHIBIT 5



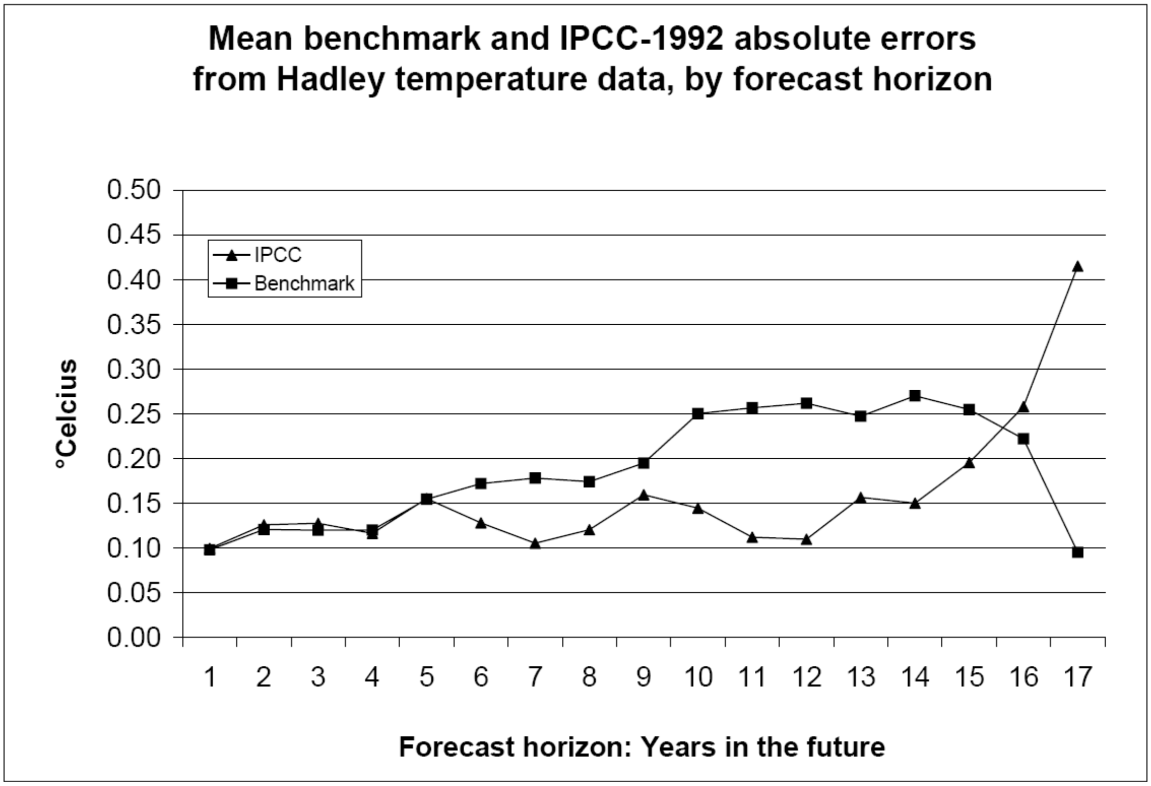
INSERT EXHIBIT 6



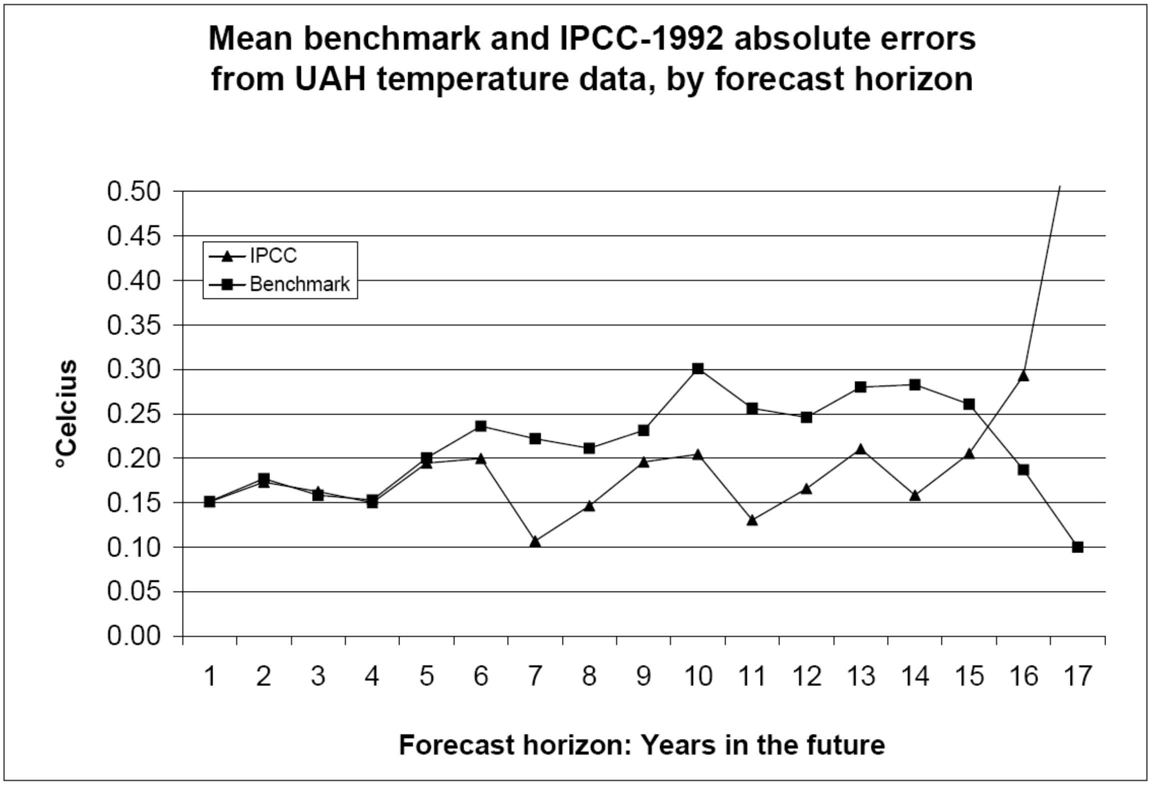
We employed successive forecasting by using each year of the Hadley data from 1991 out to 2007 in turn as the base from which to forecast from one up to 17 years ahead. We obtained a total of 136 forecasts from each of the 1992 IPCC model and our benchmark model over horizons from one to 17 years. We found that averaged across all 17 forecast horizons, the 1992 IPCC model forecast errors for the period 1992 to 2008 were 16% smaller than errors from our benchmark; the CumRAE was 0.84. The average benchmark and 1992-IPCC forecast errors for each of the 17 horizons are shown in Exhibit 7; the IPCC errors were large in the longest two horizons, as an inspection of Exhibit 5 would lead one to expect.

We repeated the successive forecasting test using UAH data. The 1992 IPCC model forecast errors for the period 1992 to 2008 were 5% smaller than errors from our benchmark (CumRAE 0.95). The series are shown in Exhibit 8. The scale is the same as for Exhibit 7 (based on the Hadley series) for ease of comparison, but this means that the 17-year-horizon IPCC error is, at 0.61°C, off the chart.

INSERT EXHIBIT 7



INSERT EXHIBIT 8



Assessed against the UAH data, the average of the mean errors for all 17 horizons was 0.215°C for rolling forecasts from the benchmark and 0.203°C for the IPCC model forecasts. The IPCC forecasts thus provided an error reduction of 0.012°C for this small sample. Such a small improvement would have no value to decisions makers. Indeed, it is hard to see how even a perfect forecast (representing an average error reduction of 0.215°C) would be useful in comparison to the already small benchmark error.

The concern of policymakers is with long-term climate forecasting, and the ex ante analysis we have described was limited to a small sample of short-horizon forecasts. To address these limitations, we used backcasting. The procedure is described in Armstrong (1985, pp. 343-345).

*Backcasts from 1974 through 1850 using 1992 IPCC model*

We used the procedure of backcasting because, as we discussed earlier, the IPCC's proposed forecasting model should be just as relevant going backwards in time from 1974 as it is going forward.

We first created a single backcast series by starting with the 1975 Hadley figure and subtracting the 1992-IPCC-model's 0.03°C from each year, starting with 1974, and repeated the process all the way back to 1851 with a backcast for 1850. Our naïve benchmark backcast was equal to the 1975 Hadley figure for all years. This process provided backcast data for each of the 125 years.

The 1992 IPCC backcast errors totaled more than ten times the benchmark errors (CumRAE 10.4). We also tested the 2007 IPCC's weaker Scenario-B trend of 0.02°C p.a. (IPCC 2007, p. 13), but it made little difference to the relative accuracy of the backcast; the 2007 IPCC errors were in total nearly seven times larger than the benchmark errors (CumRAE 6.72).

We then successively backcast by using each year from 1975 back to 1851 as the base from which to backcast from one up to 93 years back using the 1992 IPCC model and our benchmark model. This yielded a total of 7,550 backcasts covering the period 1974 to 1850 for horizons from one to 100 years.

We found that across all forecast horizons, the 1992-IPCC-model backcast errors for the period were more than seven-times greater than errors from our benchmark (CumRAE 7.23). The relative errors increased rapidly with backcast horizon. For example for horizons one-through-10 the CumRAE was 1.45, while for horizons 41-through-50 it was 6.77 and for horizons 91-through-100 it was 12.6.

## Implications for climate policy

To base public policy decisions on forecasts of global mean temperature one would have to show that changes are forecastable and that a valid evidence-based forecasting procedure would provide more accurate forecasts than those from the benchmark model. To our knowledge, this study is the first attempt to address these issues.

We did not address the issue of forecasting the net benefits or cost of any climate change that might be forecast. Here again one would need to establish a benchmark forecast, presumably a model assuming that changes in either direction would have no net effects. Researchers who have examined this issue are not in agreement on what is the optimum temperature.

Finally, success in forecasting climate change and the effects of climate change must then be followed by valid forecasts of the effects of alternative policies. And, again, one would need benchmark forecasts; presumably based on an assumption of taking no action, as that is typically the least costly. As we noted in Armstrong, Green and Soon (2008), this was overlooked in the U.S. Department of the Interior's assessment of the polar bear issue.

The problem is complex. A failure at any of one of the three stages of forecasting—temperature change, impacts of changes, and impacts of alternative policies—would imply that climate change policies have no scientific basis.

Our findings suggest that the apparently hopeless task of forecasting climate change should be abandoned.

## Conclusions

Our analyses showed that global mean temperatures are remarkably stable over policy-relevant horizons. The benchmark forecast is that the global mean temperature for each year for the rest of this century, as measured by UAH or similar, will be within 0.5°C of the 2008 figure.

There is little room for improving the accuracy of forecasts from our naïve benchmark model. In fact, it is difficult to conceive what practical benefits could be gained by obtaining forecasts. While the Hadley temperature data from thermometers shown in Exhibit 2 in retrospect appeared to drift broadly upwards over the last century or so, the longer series in Exhibit 1 shows that such trends can occur naturally over long periods. Moreover there is some concern that the upward trend might be at least in part an artifact of measurement error (e.g., heat island effects) rather than a genuine global warming. Even if one puts these reservations aside, our analysis shows that errors from our naïve benchmark forecasts would have been so small that they would not have been of concern to decision makers who relied on them. For all practical purposes, global mean temperatures are not forecastable.

Earlier research has shown that the IPCC forecasting methods violated scientific forecasting principles and IPCC forecasts should not, therefore, be used for making public policy decisions. Our findings in this paper reinforce that conclusion. We showed that a naïve no-change benchmark model produces forecasts that are sufficiently accurate for public policy decision making, and that the IPCC's forecasts are less accurate when tested against a large sample of ex ante observations.

The small sample of 17 years of IPCC 1992-model forecasts was similar in overall accuracy to the naïve benchmark forecasts. Rolling forecasts from 1992 through 2007 using the IPCC's model were only trivially more accurate than the benchmark forecasts and the mean error reduction of 0.012°C would not be useful for policy recommendations.

Climate policy is concerned with longer horizons and so our small sample of short horizon forecasts was a weak test. To address these issues we tested the relative accuracy of the IPCC forecasts using rolling backcasts over horizons of up to 100 years. We found that the IPCC backcast errors were seven times larger than those from our naïve benchmark, and the relative errors increased as the backcast horizon increased.

**Acknowledgements**

Spyros Makridakis, Malcolm Wright.

## REFERENCES

Armstrong, J.S. (1985). *Long-Range Forecasting*. New York: John Wiley.

Armstrong, J.S. (2001). *Principles of Forecasting.* Boston: Kluwer.

Armstrong, J. S., Green, K. C., & Soon, W., (2008), Polar Bear Population Forecasts: A Public-Policy Forecasting Audit, *Interfaces* (with commentary and reply), 38, 381-405.

Camp, C. D., & Tung, K.-K. (2007). Surface warming by the solar cycle as revealed by the composite mean difference projection. *Geophysical Research Letters*, 34, L14703, doi:10.1029/2007GL030207.

Carter, R. M., de Freitas, C. R., Goklany, I. M., Holland, D., & Lindzen, R. S. (2006). The Stern Review: A Dual Critique, Part I: The Science. *World Economics*, 7, 167-198.

Green, K.C., & Armstrong, J.S. (2007). Global warming: Forecasts by scientists versus scientific forecasts, *Energy & Environment*, 18, 997-1022.

Hansen, J. (2008). Tipping point: Perspective of a climatologist. In *State of the Wild 2008-2009: A Global Portrait of Wildlife, Wildlands, and Oceans*. W. Woods, Ed. Wildlife Conservation Society/Island Press, pp. 6-15.

IPCC (1990). *Climate Change: The IPCC Scientific Assessment*. Edited by J.T. Houghton, G.J. Jenkins, and J.J. Ephraums. Cambridge University Press: Cambridge, United Kingdom.

IPCC (1992). *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment*. Edited by J.T. Houghton, B.A. Callander, and S.K. S.K. Varney. Cambridge University Press: Cambridge, United Kingdom.

IPCC (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* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Lindzen, R. S. (2009). Climate Science: Is it currently designed to answer questions? Forthcoming in proceedings of *Creativity and Creative Inspiration in Mathematics, Science, and Engineering: Developing a Vision for the Future: San Marino, August 2008*.

McKibben, W. (2007). Warning on warming. *New York Review of Books*, 54, 15 March.

McKitrick, R., & Michaels, P. J. (2007). Quantifying the influence of anthropogenic surface processes and inhomogeneities on gridded global climate data. *Journal of Geophysical Research*, 112, doi:10.1029/2007JD008465.

Pielke, Jr., R. A. (2008). Climate predictions and observations. *Nature Geoscience*, 1, 206.

Soon, W. (2007) Implications of the secondary role of carbon dioxide and methane forcing in the climate change: Past, present and future. *Physical Geography*, 28, 97-125.

Soon, W. (2009). The solar Arctic connection on multidecadal to centennial timescales: Empirical evidence, mechanistic explanation, and testable consequences. *Physical Geography*, under review.