ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

international journal of forecasting

www.elsevier.com/locate/ijforecast

Validity of climate change forecasting for public policy decision making

Kesten C. Green[a,*], J. Scott Armstrong[b,1], Willie Soon[c,2]

[a] *Business and Economic Forecasting, Monash University, Vic 3800, Australia*
[b] *The Wharton School, University of Pennsylvania, 747 Huntsman, Philadelphia, PA 19104, United States*
[c] *Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, United States*

Abstract

Policymakers need to know whether prediction is possible and, if so, whether any proposed forecasting method will provide forecasts that are substantially more accurate than those from the relevant benchmark method. An inspection of global temperature data suggests that temperature is subject to irregular variations on all relevant time scales, and that variations during the late 1900s were not unusual. In such a situation, a "no change" extrapolation is an appropriate benchmark forecasting method. We used the UK Met Office Hadley Centre's annual average thermometer data from 1850 through 2007 to examine the performance of the benchmark method. The accuracy of forecasts from the benchmark is such that even perfect forecasts would be unlikely to help policymakers. For example, mean absolute errors for the 20- and 50-year horizons were 0.18 °C and 0.24 °C respectively. We nevertheless demonstrate the use of benchmarking with the example of the Intergovernmental Panel on Climate Change's 1992 linear projection of long-term warming at a rate of 0.03 °C per year. The small sample of errors from *ex ante* projections at 0.03 °C per year for 1992 through 2008 was practically indistinguishable from the benchmark errors. Validation for long-term forecasting, however, requires a much longer horizon. Again using the IPCC warming rate for our demonstration, we projected the rate successively over a period analogous to that envisaged in their scenario of exponential CO_2 growth—the years 1851 to 1975. The errors from the projections were more than seven times greater than the errors from the benchmark method. Relative errors were larger for longer forecast horizons. Our validation exercise illustrates the importance of determining whether it is possible to obtain forecasts that are more useful than those from a simple benchmark before making expensive policy decisions.

© 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Keywords: Climate model; Ex ante forecasts; Out-of-sample errors; Predictability; Public policy; Relative absolute errors; Unconditional forecasts

1. Introduction

We examine procedures that should be used to evaluate forecasts of global mean temperatures over the policy-relevant long term. In order to use forecasts to inform public policy decisions, evidence that the proposed forecasting procedure can provide *ex ante*

* Corresponding address: PO Box 10800, Wellington 6143, New Zealand. Tel.: +64 4 976 3245; fax: +64 4 976 3250
E-mail addresses: kesten@me.com (K.C. Green), armstrong@wharton.upenn.edu (J.S. Armstrong), wsoon@cfa.harvard.edu (W. Soon).
URL: http://jscottarmstrong.com (J.S. Armstrong).
[1] Tel.: +1 610 622 6480.
[2] Tel.: +1 617 495 7488.

0169-2070/$ - see front matter © 2009 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2009.05.011

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011

forecasts that are substantively more accurate than those from a simple benchmark model is needed. By *ex ante* forecasts, we mean forecasts for periods that were not taken into account when the forecasting model was developed.[3]

Benchmark errors provide a 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 could provide useful forecasts. When benchmark errors are small, it is less likely that other methods would provide improvements in accuracy that would be useful to decision makers.

2. An appropriate benchmark model

Fig. 1 displays Antarctic temperature data from the ice-core record for the 800,000 years up to 1950. The temperatures are relative to the average for the last one thousand years of the record (950 to 1950 AD), in degrees Celsius. The data show large irregular variations and no obvious trend. For such data, the no-change forecasting model is an appropriate benchmark.

3. Performance of the benchmark model

We used the Hadley (HadCRUT3) "best estimate" annual average temperature differences from 1850 to 2007 from the UK Met Office Hadley Centre[4] to examine the benchmark errors for global mean temperatures (Fig. 2[5]) over policy-relevant forecasting horizons.

3.1. Errors from the benchmark model

We used each year's mean global temperature as a forecast of each subsequent year's temperature, 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 this benchmark 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.

Fig. 3 shows that the mean absolute errors from our benchmark 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 slightly more than 0.3 °C for one-year-ahead forecasts to less than 1.0 °C for 100-year-ahead forecasts.

Overwhelmingly, the errors were no more than 0.5 °C, as shown in Fig. 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-ahead were within 1 °C of the Hadley figure. The overall maximum error from all 10,750 forecasts for all horizons was 1.08 °C (from an 87-year-ahead forecast for the year 1998).

4. Performance of the Intergovernmental Panel on Climate Change's projections

Since the benchmark model performs so well, it is hard to determine what additional benefits public policymakers would get from a better forecasting model. Governments did, however, via the United Nations, establish the IPCC to search for a better model. The IPCC projections provide an opportunity to illustrate the use of the benchmark. Our intent in this paper is not to assess what might be the true state of the world; rather, it is to illustrate proper validation by testing the IPCC projections against the benchmark model.

We used the IPCC's 1992 projection, which was an update of their 1990 projection, for our demonstration. The 1992 projection was for a linear increase of 0.03 °C per year (IPCC, 1990, p. xi; IPCC, 1992, p. 17).

The IPCC (1992) projections were based on the judgments of the authors of the IPCC report, and the process they used was not specified in such a way

[3] The ability of a model to fit time series data bears little relationship to its ability to forecast, a finding that has often puzzled researchers (Armstrong, 2001, pp. 460–462).

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

[5] Fig. 2 has been updated to include the 2008 data.

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011

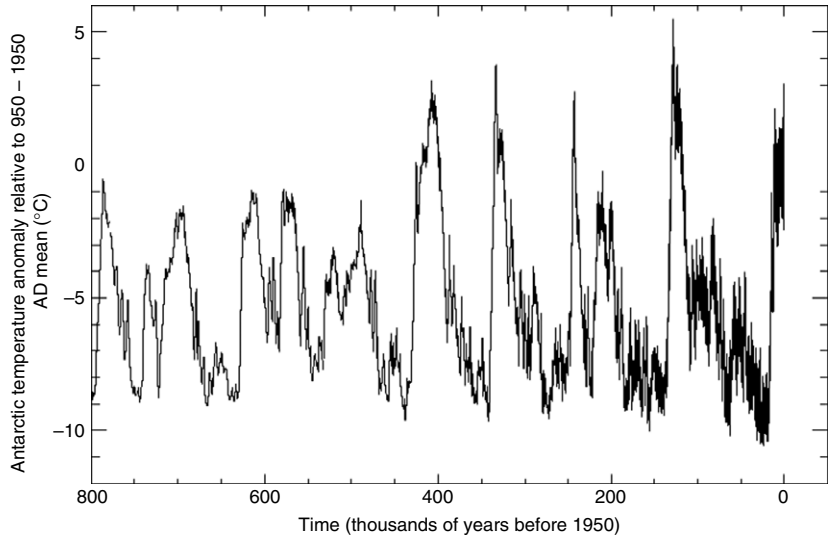

Fig. 1. 800,000-year record of Antarctic temperature change.

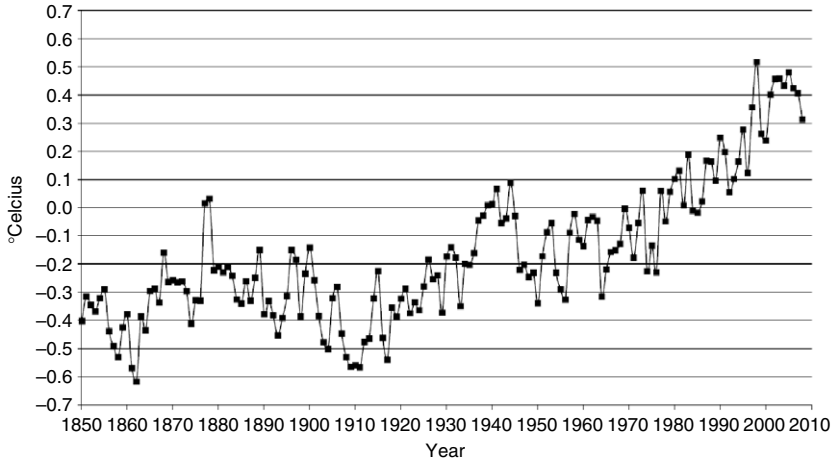

Fig. 2. Hadley annual temperature data for 1850 to 2008. Global surface temperature deviation from the 1961–1990 average.

that it would be replicable. We nevertheless used the IPCC projection because it has had a major 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 webpage on the Summit,[6] "The Earth Summit influenced all subsequent UN conferences…".

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. By using the IPCC's 1992 projection, we were able to conduct a longer *ex ante* forecasting test than if we had used projections from later IPCC reports.

4.1. 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 (1992)

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

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011

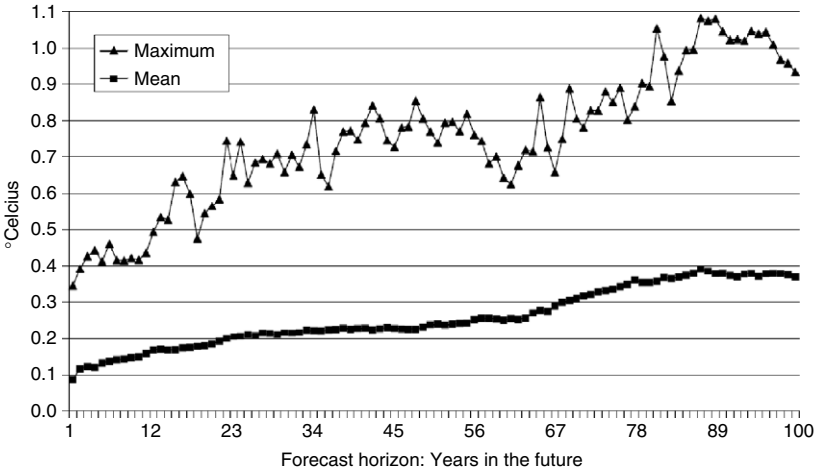

Fig. 3. Mean and maximum benchmark forecast absolute errors from the Hadley temperature data, by forecast horizon.

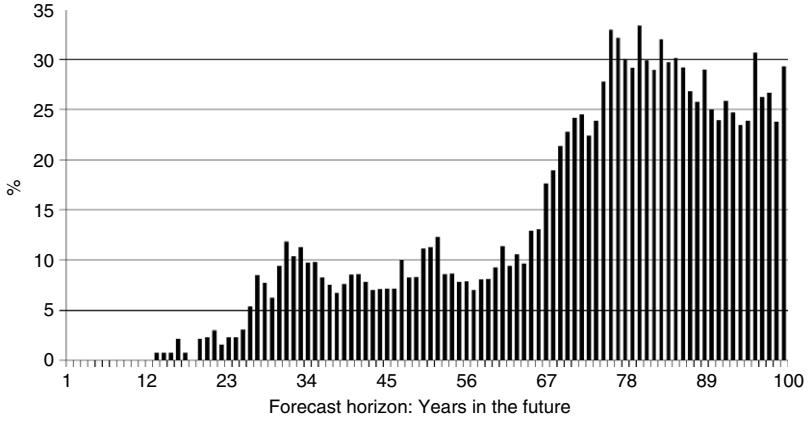

Fig. 4. Percentage of benchmark forecast absolute errors $> 0.5\,^{\circ}$C from Hadley temperature data, by forecast horizon.

projection and the Hadley figure for the same year. We then compared these IPCC projection errors with forecast errors from the benchmark model using the cumulative relative absolute error (CumRAE; Armstrong & Collopy, 1992).

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 errors of the evaluated method and the benchmark errors came to the same total, while a figure of 0.8 would indicate that the sum of the evaluated-method errors was 20% lower than the sum of benchmark errors.

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

4.2. Forecasts from 1992 through 2008 using the 1992 IPCC projected warming rate

We created an IPCC projection series from 1992 to 2008 by starting with the 1991 Hadley figure and

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011

adding 0.03 °C per year. It was also possible to test the IPCC projected warming rate against the University of Alabama at Huntsville's (UAH) data on global near surface temperature measured from satellites using microwave sounding units. These data are available from 1979 onwards. To do this, we created another projection series starting with the 1991 UAH figure.

Benchmark forecasts for the two series were based on the 1991 Hadley and UAH temperatures, respectively, for all years. This process, including estimates for 2008 from both sources, gave us two small samples of 17 years of out-of-sample forecasts. When tested against Hadley measures, 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.

We also employed successive updating by using each year of the Hadley data in turn, from 1991 to 2007, as the base from which to forecast from one to 17 years ahead. We obtained a total of 136 forecasts from each of the 1992 IPCC projected warming rate and the benchmark model over horizons from one to 17 years. We found that, averaged across all 17 forecast horizons, the 1992 IPCC projected warming rate errors for the period 1992 to 2008 were 16% smaller than forecast errors from our benchmark, as the CumRAE was 0.84.

We repeated the successive forecasting test using UAH data. The 1992 IPCC projected warming rate errors for the period 1992 to 2008 were 5% smaller than forecast errors from our benchmark (CumRAE = 0.95).

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 model and 0.203 °C for the IPCC projected warming rate. The IPCC projections thus provided an error reduction of 0.012 °C for this small sample of short-horizon forecasts. The difference of 0.012 °C is too small to be of any practical interest.

Policymakers are concerned with long-term climate forecasting, and the *ex ante* analysis we have described was limited to a small sample of short-horizon projections. To address this limitation, we calculated rolling projections from 1851 to illustrate a proper validation procedure.

4.3. Forecasts from 1851 through 1975 using the 1992 IPCC projected warming rate

Dangerous manmade global warming became an issue of public concern after NASA scientist James Hansen testified on the subject to the US Congress on June 23, 1988 (McKibben, 2007), after a 13-year period from 1975 over which global temperature estimates were up more than they were down. The IPCC (2007) authors explained, 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). There have even been claims that human activity has been causing global warming for at least 5000 years (Bergquist, 2008).

It is not unreasonable, then, to suppose, for the purposes of our validation illustration, that scientists in 1850 had noticed that the increasing industrialization of the world was resulting in an exponential growth in "greenhouse gases", and projected that this would lead to global warming of 0.03 °C per year.

We used the Hadley data from the beginning of the series in 1850 through to 1975 to illustrate the testing procedure. The period is not strictly out-of-sample, however, in 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 is some positive trend, so the benchmark is disadvantaged for the period under consideration. As is shown in Fig. 1, the variations in the longer temperature series suggest there is no assurance the irregular trend observed in retrospect will continue in the future.

We first created a single forecast series by adding the 1992 IPCC projected warming rate of 0.03 °C to the previous year's figure, starting with the 1850 Hadley figure, and repeating the process for each year through to 1975. Our benchmark forecast was equal to the 1850 Hadley figure for all years. This process provided forecast data for each of the 125 years. The IPCC warming-rate projection errors totaled more than ten times the benchmark errors (CumRAE = 10.1).

We then successively used each year from 1850 to 1974 as the base from which to forecast from one

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011

up to 100 years ahead using the 1992 IPCC projected warming rate and the benchmark model. This yielded a total of 7550 forecasts covering the period 1851 to 1975. Across all horizons, the projection errors for the period were more than seven times greater than errors from our benchmark (CumRAE = 7.67). The relative errors increased rapidly with the horizon. For example, for horizons one through ten, 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.

Thus, even though this was a period during which warming occurred, the IPCC projection rate would have produced forecast errors that were more than 12 times those from the benchmark model.

5. Discussion

We have illustrated how to validate a forecast. There are other reasonable validation tests for global mean temperatures. For example, one reviewer argued that the relevant forecasts for climate change are for decades or longer periods. For decadal forecasts, the appropriate benchmark forecast is that the decades ahead will be the same as the decade just gone. The mean absolute error of a rolling one-decade-ahead benchmark forecast, calculated using the entire Hadley series from 1850 to 2007, was 0.104 °C. The Mean Absolute Error (MAE) for five decades ahead was 0.198 °C, and for 10 decades ahead was 0.345 °C. These decadal benchmark errors are smaller than the annual benchmark errors.

Validation tests should properly be conducted on forecasts from evidence-based forecasting procedures. The models should be clearly specified, fully-disclosed, and replicable. The conditions under which the forecasts apply should be described.

Speculation is not sufficient for forecasting. The belief that "things have changed" and that the future cannot be judged by the past is common, but invalid. The 1980 bet between Julian Simon and Paul Ehrlich on the 1990 price of resources was a high-profile example. Ehrlich espoused the Malthusian view that the human population's demands had outstripped, or soon would outstrip, the resources of the Earth. Simon's position was that real resource prices had fallen over human history, and that there were good reasons why this was so; the fundamental reason being human ingenuity. It was therefore a mistake, Simon maintained, to extrapolate recent price increases. Ehrlich dictated the terms of the bet: a ten-year period and the five commodity metals copper, chromium, nickel, tin, and tungsten. The metals were selected with the help of energy and resource experts John Harte and John P. Holdren. All five commodities fell in price over the ten-year period, and Simon won the bet (Tierney, 1990).

To base public policy decisions on forecasts of global mean temperature, one would have to show that changes are forecastable over policy-relevant horizons, and that a valid evidence-based forecasting procedure would provide usefully more accurate forecasts than those from the "no change" benchmark model.

We have not addressed the issue of forecasting the net benefit or cost of any climate change that might be predicted. 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 as to what the optimum temperature is.

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.

The problem is a complex one. A failure at any 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.

6. Conclusions

Global mean temperatures have been remarkably stable over policy-relevant horizons. The benchmark forecast is that the global mean temperature for each year for the rest of this century will be within 0.5 °C of the 2008 figure.

There is little room for improving the accuracy of forecasts from our benchmark model. In fact, it is questionable whether practical benefits could be gained by obtaining perfect forecasts. While the Hadley temperature data in Fig. 2 drifts upwards over the last century or so, the longer series in Fig. 1 shows that such trends can occur naturally over long periods before reversing. Moreover, there is some

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011

concern that the upward trend observed over the last century and half might be at least in part an artifact of measurement errors rather than a genuine global warming (McKitrick & Michaels, 2007). Even if one accepts the Hadley data as a fair representation of temperature history, our analysis shows that errors from the benchmark forecasts would have been so small that decision makers who had assumed that temperatures would not change would have had no reason for regret.

Acknowledgements

We thank the nine people who reviewed the paper for us at different stages of its development and the two anonymous reviewers for their many helpful comments and suggestions. We also thank Max Feldman and Michael Guth for their useful suggestions on the writing.

References

Armstrong, J. S. (2001). Evaluating forecasting models. In J. S. Armstrong (Ed.), *Principles of forecasting* (pp. 443–472). Boston: Kluwer Academic Publishers.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*, 69–80.

Bergquist, L. (2008). Humans started causing global warming 5000 years ago, UW study says. *Journal Sentinel*, posted 17 December. http://www.jsonline.com/news/education/36279759.html.

IPCC (1990). In J. T. Houghton, G. J. Jenkins, & J. J. Ephraums (Eds.), *Climate change: The IPCC scientific assessment*. Cambridge, United Kingdom: Cambridge University Press.

IPCC (1992). In J. T. Houghton, B. A. Callander, & S. K. Varney (Eds.), *Climate change 1992: The supplementary report to the IPCC scientific assessment*. Cambridge, United Kingdom: Cambridge University Press.

IPCC (2007). Summary for policymakers. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, & H. L. Miller (Eds.), *Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge, UK, New York, NY, USA: Cambridge University Press.

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.

Tierney, J. (1990). Betting the planet. *New York Times*, December 2.

Please cite this article in press as: Green, K. C., et al. Validity of climate change forecasting for public policy decision making. International Journal of Forecasting (2009), doi:10.1016/j.ijforecast.2009.05.011