Lancaster University Management School Working Paper 2010/010

Validation in models of climate change and forecasting accuracy

Robert Fildes and Nikolaos Kourentzes

The Department of Management Science
Lancaster University Management School
Lancaster LA1 4YX
UK

The LUMS Working Papers series can be accessed at http://www.lums.lancs.ac.uk/publications/
LUMS home page: http://www.lums.lancs.ac.uk/

Validation in models of climate change and forecasting accuracy

Robert Fildes and Nikolaos Kourentzes, Lancaster Centre for Forecasting
Lancaster University Department of Management Science
Email: R.Fildes@Lancaster.ac.uk

Forecasting researchers, with few exceptions, have ignored the major forecasting controversy facing the world in the early 21st Century: namely, whether and by how much the earth is warming; and the role of climate modelling in reaching any conclusions on this challenging topic. In contrast, scientists from climatologists, through hydrologists to fluid dynamicists, have engaged in this modelling and forecasting controversy. In this discussion paper, we first describe briefly the atmospheric-ocean general circulation models (AOGCM) used in most climate forecasting, in particular by the Intergovernmental Panel on Climate Change (IPCC). This discussion paper takes a forecaster's perspective in a review of established principles for the validation of such large-scale simulation models. One key principle is that such models should reproduce the 'stylised facts' or 'dominant modes' of dynamic behaviour that characterize key model outputs: here taken as the aggregate annual changes in world and regional temperatures. By developing various time series models and input-output dynamic models of atmospheric carbon dioxide and temperature that capture current trends, we compare the results with dynamic forecasts produced by one well-established AOGCM model, the Hadley Centre's HadCM3. Time series models are shown to perform strongly and by using encompassing tests, structural deficiencies are identified in the AOGCM model and its corresponding forecasts. The paper concludes with some implications for climate modellers when producing decade-ahead forecasts from global climate models. If forecasting accuracy is the focus, methods that combine standard time series methods with the structure of a GCM should be used. This has implications for the effectiveness of control policies, focussed on carbon dioxide emissions alone..

Keywords: validation; long ranger forecasting; simulation models; Global circulation models; neural nets

February 2010

Validation in models of climate change and forecasting accuracy

Robert Fildes and Nikolaos Kourentzes, Lancaster Centre for Forecasting.

1. Introduction

Of all the areas of forecasting that have succeeded in gaining public attention, current forecasts of global warming and the effects of human activity on the climate must surely be in first place. Even before the Kyoto treaty of 1997 there has been an emerging scientific consensus identified with the Intergovernmental Panel on Climate Change (IPCC). By the time of its fourth report in 2007, there were few scientists working in the field who did not accept two central tenets from the IPCC's work: that the earth was warming and that some part of that was due to human activity. Nevertheless, there had long been powerful counter-voices, both political (e.g. the last Australian Prime Minister John Howard and ex-President George W. Bush) and scientific (embodied in the work of the Heartland Institute and its publications (Singer and Idso, 2009) and conferences, supported by the arguments of a number of eminent scientists, some of whom research in the field (see Lindzen, 2009). While various political and industrial biases are apparent in many of the contrarian criticisms, the vehemence with which sceptics have assailed the IPCC raises questions as to why the 4th Report is viewed by many as providing inadequate evidence. The aims of this discussion paper are to review the various criteria used in appraising the validity of climate models, in particular the role of forecasting accuracy comparisons and to provide a forecasting perspective on this important debate that has so far been dominated by climate modellers. We focus on decadal long forecasts (10 to 20 years ahead). Such forecasts have many policy-relevant implications, from land-use and infrastructure planning to insurance, and climatologists have shown increasing interest in this "new field of study" (Meehl and al., 2009). Decadal forecasts also provide sufficient data history for standard forecasting approaches to be used.

In section two of this discussion paper we first set out accepted principles underlying the notion of a 'valid forecasting model', particularly as they apply to complex mathematical models such as those used in climate modelling. The evaluation of such models is necessarily multi-faceted, but here we pay particular attention to the role of forecasting benchmarks and forecast encompassing, an aspect neglected by climate modellers generally as well as in the IPCC discussion of the evaluation of climate models in chapter 8 of the Fourth Report (Randall, et al., 2007). In section 3 we provide empirical evidence on the forecasting accuracy 10 and 20 years ahead for global average temperatures using benchmark univariate and multivariate forecasting methods. In particular, a multivariate neural net is used to produce the forecasts. We examine the effect of including CO_2 emissions on forecasting performance. These results are contrasted with those produced by Smith et al. (2007) using one of the Hadley Centre's models, HadCM3 and its decadal predictive variant, DePreSys. By considering forecast combining and encompassing it is shown that the trends captured in the time series models contain information not yet included in the HadCM3 forecasts. Section 3 also considers some further aspects of model validation; in particular disaggregate forecasts of local temperatures.

While our results show evidence of global warming from an ex ante forecasting perspective, there is only limited evidence of an ex ante predictive relationship between annual emissions, atmospheric CO_2 and the 10 and 20-year-ahead global annual average temperature. The last section reflects on the importance of comparative forecasting accuracy, offering recommendations to the climate-change scientific community on the benefits of adopting a multidisciplinary modelling perspective that incorporates the lessons learnt from forecasting research.

2. Simulation Model Validation in Longer-term Forecasting

The models at the heart of the IPCC report, while differing in detail, are all examples of Coupled Atmospheric-Ocean General Circulation Models (AOGCM)[1]. A brief summary of their basis is as follows. They are systems of partial differential equations based on the basic laws of physics, fluid motion, and chemistry. To 'run' a model, scientists divide the planet into a 3-dimensional grid plus time, apply the basic flow equations to calculate winds, heat transfer, radiation, relative humidity, and surface hydrology within each grid and evaluate interactions with neighboring points. The outputs includes temperature and precipitation estimates across the grid as well as many other variables, and these are averaged to produce such publicly high profile outputs as 'average global temperature'. Figure 1 shows a stylised representation of such models.

Initial conditions and parameters must be set to solve numerically the partial differential equations at the heart of the model. The initial conditions are fixed depending on the starting

[1] In addition, smaller scale models focusing on aspects of the world's climate are also used. The high level aggregate forecasts are produced from the AOGCMs.

point of the runs, often many hundreds of years ago. At that distance in the past, the observations are limited (from such measures as ice core) and therefore the starting values are based on arbitrary pre-industrial states (Meehl at al, 2009). The parameters in the model are selected based on “simplified physical models”. Alternatively, they may be ‘tuned’ (estimation or calibration in forecasting terminology), whilst remaining compatible with prior information and established physical relationships, so that the outputs of the simulation ‘fit’ particular observed outputs and spatial relationships (data assimilated in climate modeling terms). The aim is to provide a ‘best’ estimate of the true state of the world climate system and corresponding prediction equations for simulating recent climate history and for forecasting. The start-up runs typically drift so that by the time data are more readily available, there is often a discrepancy with the simulated output. Further tuning is used to ensure the model is back on track (e.g. “to calculate the ocean-heat flux convergence field”, Stainforth et al, 2002). In addition, from 1850 approximately, observed data on ‘forcing’, exogenous variables in statistical terminology, such as CO_2 emissions are included. Because of the model complexity, the computer costs of optimization of these steps is currently prohibitive. Even if it were feasible, given the large number of degrees of freedom and the limited observations judgment is necessarily used. Thus, a major part of the model building is judgmental (Stainforth et al, 2007).

With the model ‘on-track’, the prediction equations roll out the current system states over time to deliver forecasts of many variables across time and space, of which there are a number that are regarded as key to model performance. A distinction has been drawn between long-term (100+ years ahead) prediction and decade-ahead predictions. In the former task “the climate models are assumed to lose all memory of their initial conditions” and thus, current observations are not used to ground (or ‘assimilate’) the model in the data. Note that the observed data form only a small sub-set of the GCM’s output. For decade-ahead forecast horizons the recent conditions matter so that to produce plausible forecasts, the models must be calibrated to the current observations. For the IPCC forecasts this has not been done since their perspective is primarily on the longer term. Recently, various modelling exercises have focussed, for reasons we have already explained, on decadal prediction (Smith, 2007; Haines et al, 2009; Meehl et al, 2009). The forecasts from the GCMs use as their initial values the observations at the forecast origins, as we explain in greater detail in section 3.

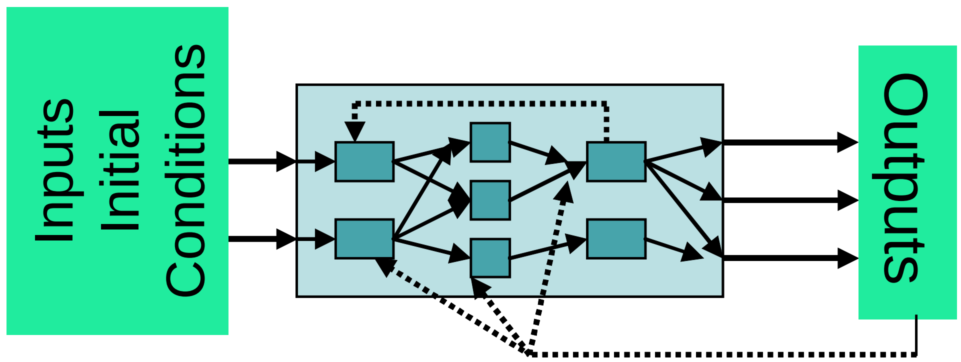

Figure 1: Stylised representation of Global Circulation Climate Models (GCMs)

The prevalent research strategy in the climate-modelling community has been characterised by Knutti (2008), himself a climate modeller, as "take the most comprehensive model ..., run a few simulations ...at the highest resolution possible and then struggle to make sense of the results". The aim is to produce models as "realistic as possible" (Beven, 2002). However, various models of sub-systems (e.g. Earth Systems Models of Intermediate Complexity (EMICs) have been constructed delivering simpler models that are more manageable. See Claussen et al., (2002) for a discussion of a "spectrum of climate system models" which differ as to their complexity though with AOGCMs at the extreme.

There is feedback between the outputs and pre-cursor variables with varying, often long lags, and non-linearities, for example Young and Jarvis (2002) show a non-linear feedback relationship between emissions, an input, an intermediate variable, atmospheric CO_2 that is affected by a key output temperature. Lags of up to 1000 years are expected within the system because of factors such as the slow warming (or cooling) of the deep seas.

In considering the validity of AOGCMs (or more generally, environmental simulation models) various authors have examined where errors in a model's predictions may arise, for example, Kennedy and O'Hagan (2001), Beven (2002) and Stainforth et al. (2007). The characterisation of model error that follows is compatible with their views. Uncertainty in the conditional model-based forecasts arises from a number of sources:

(i) the initial conditions

- To solve the model and produce predictions, the partial differential equations need to be initialised. The choice is arbitrary but nevertheless affects the results. One response of General Circulation Modellers is run the model for a small number of initial states.

This results in a distribution of outcomes (see e.g. Stainforth et al., 2007, Fig. 1). The final forecasts are based on an average of the results that may exclude 'counter-intuitive' realisations (Beven, 2002).

(ii) various parameters that are not determined by the physics of the models but are approximate estimates

- In fact it is rare that model parameters are uniquely determined from theoretical considerations. They will depend on many factors including the specific location where they are applied (Beven, 2002: section 3). Nor does the problem disappear with increased disaggregation, with Beven arguing it may make matters worse.

The parameters in a GCM are sometimes 'tuned', but rarely optimally estimated. When a set of parameters is estimated they are likely to suffer from the standard problem of multicollinearity. A key point to note is that possible non-linear effects, e.g the CO_2 absorption capacity of a forest at levels of atmospheric $C0_2$ twice that currently observed cannot be known or reliably estimated.

(iii) uncertainty arising from model mis-specification

- For example, in the current generation of AOGCMs certain potentially important processes such as cloud effects and water vapour formation are still poorly understood. A second example is how vegetation is modelled. Aggregation over time and space also leads to mis-specification. A necessary consequence of parameter uncertainty and specification uncertainty is that none of the set of models (in model space, Beven's 2002 terminology) that represent the global climate, can be rejected in a "relaxed form of Popperian falsification" (Beven, 2002). However, they all fail in some important attributes. The common view, Knutti (2008) claims, is that they all offer "credible approximations to the descriptions of the climate system given our limited understanding".

(iv) randomness

- With stochastic models, this is always an important source of uncertainty. Even if the nature of the models is essentially deterministic (as with GCMs), this still remains potentially important since the paths taken are likely to be state dependent. As a consequence, small (even localised) discrepancies may accumulate.

(v) uncertainty in the data

- There remains considerable controversy as to the choice of measure for the key variable, temperature, whether at an aggregate level or even at more local levels where changes in the local environments such as increased urbanisation provide the basis for a critique of the raw data (Pielke, et al., 2007).

(vi) numerical and coding errors

- In the solution to the system equations, unavoidable numerical errors may occur as well as coding errors.

If unconditional forecasts are required, additional uncertainty arises from the unknown future levels of the forcing inputs such as volcanic eruptions and CO_2 emissions. (Confusingly these are sometimes referred to in the climate literature as ‘boundary conditions’.)

Various approaches to mitigate these uncertainties have been proposed. Ensemble methods provide a combined set of predictions based on different initial conditions (Hagedorn et al., 2005). In addition, some aspects of the specification uncertainty are alleviated through multi-model averaging. The results of comparing the benefits of the two approaches to a alleviating uncertainty in within-year seasonal forecasting is that there is more uncertainty arising from the various model specifications than from the initial conditions (Hagedorn, et al., 2005). The similarities with the ‘combining’ literature that long predates this research have not been noted in the discussions on climate.

There is a current debate on appropriate methods of model averaging (Lopez, et al., 2006). A Bayesian approach (Tebaldi, et al., 2005) weights models depending on their conformity with current observations. More controversially, the weighting associated with an individual model is related to how closely its forecasts converge to the ensemble mean (based on the unrealistic assumption of the models being independent drawings from a superpopulation of AOGCMs). This leads to a probability density functions, either uni- or multimodal, the latter being the result of the models disagreeing. Substantially different results arise from these different methods. As yet there is no reason to believe the conclusion of this debate will depart from that in the forecasting literature: recommending a simple or trimmed average for the most accurate point forecast (Jose and Winkler, 2008). The range of forecasts from a selected group of GCMs or the estimated probability density function of the ensemble offer an understanding of the uncertainty in these ensemble forecasts. However, “there is no reason to expect these distributions to relate to the probability of real-world behaviour” (Stainforth et al., 2007) since the modelling groups and their forecasts are interdependent, sharing both a common modelling paradigm and methods, data and computer hardware. Counterintuitive forecasts that do not fit with the consensus are given low weight (as in the Bayesian combination) or omitted (if, for example, a new ice age is foreseen, Beven, 2002).

The effects of uncertainty in the forcing variables is primarily dealt with through the use of policy scenarios that aim to encompass the range of outcomes so as to guide policy and decision making (Dessai and Hulme, 2008). This approach will usually leave out possible future effects of events such as volcanic eruptions (e.g. Mt. Pinatubo eruption in 1991). These can be simulated through stochastic interventions to give an estimated distribution of future outcomes, conditional on the particular emissions scenario.

In the forecasting literature the various methods for estimating the future error distribution (see Chatfield, 2001) are all (often necessary) substitutes for observing the error distribution directly. In general, none of the model-based approaches substitute for observing the error distribution directly and the model-based estimates of the predictive density function (and prediction intervals) are unlikely to be any better calibrated in climate forecasting than in other applications (Stainforth et al., 2007). This has been recognized in principle by the IPCC although as Pielke (2008) points out, there is a need to be clear about the exact variables used in the conditional predictions and their measurement. However, there are few studies that present error distributions, in part because of the computational complexity of GCMs.

For long horizons (100+ years) climate modellers have tended to dismiss the prospect of estimating the conditional error distribution arguing that models of the effects of slower physical processes such as the carbon cycle are reliant on proxy data (e.g. ice records) and these have been used in the model construction, effectively rendering the comparison between the model forecasts and the observations an ‘in-sample’ test, which can be no more than confirmatory (Stainforth et al., 2007).

In summary, while the match between model predictions with their associated prediction intervals is recognized by all the authors we have referred to as a key criterion for appraising the different GCMs, few if any studies have formally examined their comparative accuracy record, which is at the heart of forecasting research.

2.1. Validation in Long Term Forecasting

What distinguishes decadal forecasting from its shorter-horizon relative and do any of the differences raise additional validation concerns? Curiously, the *Principles of Forecasting* (Armstrong, 2001), which aims to cover all aspects of forecasting, gives no special attention to the topic, regarding the approaches covered within as definitive and complete. Armstrong's earlier book with an apparent focus on the longer term (*Long Range Forecasting*, (Armstrong, 1985), points out the difficulty of a clear definition but suggests that what distinguishes the two is the prospect of large environmental change. Certainly, we have already seen dramatic change in forcing variable of CO_2 emissions in the past 150 years, now at levels not seen for thousands of years (if ever) with scenarios predicting a doubling over the next 50 years [2] leading to a rise of a further 2.0 - 5.4 degrees in the high-emissions IPCC scenario (A2[3]) in this century. Thus, the condition of dramatic exogenous environmental change is expected. Therefore, any forecasting model designed to link CO_2 emissions (or any other induced forcings such as changed land use) with temperature change must aim to establish a robust relationship between the two in this future, as yet unobserved, world, not just in the past.

Oreskes (1994) marshalling the logic of the philosophy of science, has argued that such open system models as the AOGCMs cannot be verified (although elements of a model such as its numerical accuracy may be), nor can they be validated in the strongest sense of the word, that is implying the veracity of the model under review. While some climate modellers have perhaps been sucked into the view that a valid model should realistically represent the 'real' system in depth and detail, forecasting researchers have taken a more comparative view of validity. To them, the validation of a model has a weaker meaning and has come to embrace two features (i) 'data congruence', whereby there are no systematic errors in the difference between what has been observed and the forecasts, and (ii) forecast encompassing, that is, the model produces more accurate forecasts than alternative forecasting models. Forecasting models (like all simulation models) are seen as only temporarily valid, designed for particular uses and users, and subject to repeated confrontations with the accumulating data (Kleindorffer et al, 1998).

[2] http://www.climatescience.gov/Library/sap/sap3-2/final-report/sap3-2-final-report-ch2.pdf, page 21, figure 2.1 and page 24

[3] http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-spm.pdf

But long-range forecasts from AOGCMs for longer policy relevant time spans, when there is considerable natural variability in the system as well as apparent non-stationarity, have not provided the necessary historic record, which would deliver supporting evidence on their accuracy. Some researchers have regarded this as conceptually impossible since waiting decades or more until the predictions are realised (and the models rerun to include various forcings such as actual emissions) is hardly a policy-relevant solution. But retroactive evaluations are the common currency of forecasting-model evaluations. Although the climate model parameters have, as noted above, been calibrated on data potentially used in the evaluation, that does not annul the utility of making the comparisons. In fact this should of course benefit the GCM results. One additional key constraint in decadal forecasts or longer is the computational requirements of running such large models and this has undoubtedly limited researchers' abilities and willingness to produce a simulated historical record.

In summary, the claim that as "realistic [a model] as possible" will necessarily produce more accurate forecasts has long been falsified within forecasting research, witness the various discussions of macroeconomic forecasting accuracy critiquing large macro models, and more recently the acceptance that small (often simple) models are the most effective (Granger and Jeon, 2003). Thus, with the focus on forecasting accuracy and its policy implications, the requirement for valid models (and forecasts) requires the construction of an accuracy record.

A contrary case can be made about the value of such a historical record in model evaluation as we discuss below. The key objection arises from the expected lack of stationarity when the models are used outside their estimation domain. Thus, the novel issue in model validation for decadal climate forecasting (or longer) using GCMs is the need to marshal supporting evidence that the models are useful for forecasting in the extended domain of increasingly high levels of atmospheric CO_2.

2.1.1 Defining the Problem Context

"All models are incorrect but some are useful.[4]" Any meaningful evaluation must specify (i) the key variables(s) of interest –such as annual average global temperature or more localised variables, (ii) a decision relevant time horizon, and (iii) the information set to be used in constructing the forecasts.

[4] Usually attributed to George Box.

With regard to specifying the variable(s) of interest and the forecast horizon, while substantial attention has been given to the aggregate forecasts, particularly of temperature, the AOGCM forecasts are highly disaggregate and their corresponding local forecasts of temperature, precipitation and extreme events have been extensively publicized and their implications for policy discussed. Thus, the disaggregate forecasts are of interest in their own right. The time horizon over which the climate models are believed to be useful in society is typically unspecified but goes from a decade to centuries ahead. In particular, they are not intended as short-term forecasting tools although in the IPCC report Randall et al (2007) take the contrasting view that “climate models are being subjected to more comprehensive tests including ... evaluations of forecasts on time scales from days to a year”. As we argued in the previous paragraphs, models accurate in the short term are not necessarily suitable for longer term forecasting (and of course, vice versa). As a consequence it is necessary to focus on a policy relevant horizon and here we have chosen a 10-20 year horizon, short by climate modelling perspective. It is, however, relevant to infrastructure upgrades, energy policy, insurance, etc. and has, as noted, become the focus of some climate modelling research (Meehl et al, 2009).

The third characteristic, the information set, is only relevant here when considering the evaluation of forecasts, where there has been some confusion over the distinction between conditional ex post evaluations (based on realised values of emissions) and unconditional ex ante forecasts (Trenberth, 2007). Since the focus of this article is the validity of the models for decadal forecasting, CO_2 emissions can be regarded as known, at least in any *ex post* evaluation. Other potential explanatory variables such as land use can be treated similarly. Unpredictable events, such as volcanic eruptions, can be treated as part of the noise and output can be tested for robustness to such cataclysmic and unpredictable events as the Mt. Pinatubo eruption. Whether forecasting with GCMs or time series models, such events can be included as part of the information base in the in-sample modelling.

A fourth feature of the problem context requires a little more discussion: who are the intended users/ consumers of the forecasts? Little (1970), as part of his influential discussion of model building, argues that models if they are to be valuable to their users should be 1. Complete on ‘important’ dimensions, 2. Comprehensible to the stakeholders, 3. Robust, and 4. Controllable, i.e. the user should be able “to set inputs to get almost any outputs”. Various modellers concerned with environmental policy have also examined the role of models. For example, Pielke (2003) proposes guidelines that support and extend Little, in particular emphasising the importance of clarity as to the uncertainties in the model and forecasts. Since we are focussing on validation within the scientific community AOGCMs achieve the first criterion (though there are still recognized omissions from the models). There has been less attention given to the remaining criteria. With such a wide range of stakeholders the IPCC have chosen to present their models for expert audiences and popularised their dramatic consequences through, for example, their 'Summary for Policy Makers'. Issues such as the robustness and controllability of the models have been kept in the hands of the model developers with the ultimate users (governmental policy makers and their populations) at a distance. Although in principle the models are comprehensible, their completeness (and complexity) means that there has been relatively little experimentation aimed to test the sensitivity of functional forms, parameterisations, or initial conditions.. However, the model comparisons being carried out in various programmes such as project GCEP (Grid for Coupled Ensemble Prediction); Haines et al. (2009) aim to overcome some of these limitations to "explore predictability".

2.1.2 Forecast (Output) Validation

In forecasting, as in science more generally, the primary criterion for a good model is its ability to predict the key variable(s) from pre-specified information. An early example of neglecting forecast validation in global modelling was in the 'Limits to Growth' system dynamics simulation model of the world (Meadows, et al., 1972) which whilst much more aggregate than the current generation of AOGCMs, included additional variables measuring population, technology and economy as well as environmental variables. Whilst aimed primarily as a policy tool, the 'Limits' authors inevitably slipped back into forecasts (conditional on various policies). No attempt was made in this early world modelling exercise to demonstrate it had any forecasting abilities when compared to alternative methods.

As part of the early debate on economic model building Friedman (1953) had elevated predictive ability above any other in his requirements of a useful economic model, arguing that too much weight (in model building) is given to the "realism of assumptions". Following Friedman (and many others) AOGCMs, therefore, should be evaluated through comparing their out-of-sample forecasts, conditional on using known values of various explanatory (forcing) variables and assumed policy-determined variables such as CO_2 emissions. The resulting forecasts can then be compared with the 'future' observations. (Other forcing variables such as volcanic emissions could be treated either as known or unknown depending on the purpose of the model evaluation.) If one model is to be preferred (on this criterion) to another then the observed errors on past data should be smaller (on the relevant measures, e.g MAPE, RMSE, turning point predictions). A fundamental contribution of forecasting research is to emphasize the requirement for a method (or forecasting process) to demonstrate its superiority by beating some plausible competitor benchmark. In so far as researchers know how to select a good forecasting method ex ante, perhaps the primary requirement is that it has been shown to work in past circumstances similar to those expected to apply in the future, outperforming alternatives, in particular a benchmark (Armstrong and Fildes, 2006). Of course it is expected that in small-samples, noise may well overwhelm the signal (in the GCMs deriving from increasing CO_2 emissions) and therefore a large sample of forecasts would need to be considered.

A number of researchers have criticised the IPCC models and forecasts[5] for their failure to provide evidence of predictive accuracy despite the IPCC's strong claims (Green and Armstrong, 2007; Pielke, 2008). At the heart of this argument is the need for the IPCC and GCM builders to apply rigorous standards of forecast evaluation to the IPCC forecasts of temperature change and other key variables. Since the forecasts from these climate models aim to induce novel policies (potentially expensive, see for example Stern, 2007), the importance of the IPCC models delivering forecasts, more accurate than competing alternatives with reliable prediction intervals cannot be overestimated. Localised forecasts derived from the AOGCMs need to be subjected to the same tests since policies will typically be implemented locally (see for example, Koutsoyiannis et al., 2008, who evaluate local temperature and precipitation forecasts at 8 locations and our discussion of the same issue in section 3.3 of this paper).

Where there are multiple outputs from a simulation model (as with AOGCMs) and no single output is elevated above the others, indices need to be constructed that take dependencies into account. (See Reichler and Kim (2008) or, within the forecasting literature, Clements and Hendry 1995).

The forcing (exogenous) variables are measured with error and features like a major volcanic eruption may have produce large errors in some models (perhaps because of dynamic effects)

[5] We use 'IPCC forecasts' as short-hand for the simulated forecasts from AOGCM models, conditional on selected scenarios, produced by various of the modelling agencies and discussed in the IPCC report.

that are not reproduced in others. This reinforces the need for robust error measures and rolling origin simulated errors (Fildes, 1992).

We conclude that the specific features of the evaluation of climate simulation models' output do not pose any fundamental issues that earlier discussions of forecasting evaluation have not considered. However, the size of these models apparently discourages the obvious resolution to this problem; fix a starting date where the exogenous variables are regarded as reliably measured (within some range), 'tune' the model to match the in-sample data and calculating the out-of-sample rolling origin forecast errors[6]. Instead, even large-scale comparisons such as that of the Program for Climate Model Diagnosis and Intercomparison (PCMDI) content themselves with short-term, primarily qualitative comparisons, such as model stability, variability of model output compared with the observed and consistency with observations, most often presented graphically (Phillips, et al., 2006). Smith et al. (2007) have attempted to overcome these limitations using a version of HadCM3, DePreSys (Decadal Climate Prediction System) which "takes into account the observed state of the atmosphere and ocean in order to predict internal variability". Thus, Smith et al. (2007) and others have demonstrated that exercises in forecast validation are in principle practical.

In summary, there is an increased recognition within the climate modelling community of the importance of decadal prediction with a consequence that there needs to be an increased emphasis on data assimilation methods to initialise the forecasts if effective forecasts are to be produced on a decadal time scale (http://www.clivar.org/organization/decadal/decadal.php)

2.1.3 Stylised Facts

A second aspect of validating a forecasting model comes into play here, the need for models capable of capturing the stylised facts of climate fluctuations. The term 'stylised fact' here is used conventionally[7] to mean a simplified characterisation of an empirical finding. Here the AOGCMs aim to simulate various stylised facts in the current record and potentially the more distant past. For example, such stylised facts include the changing temperature trend over the last century, the effects of major volcanic eruptions, the cyclical effects of El Nino and the Southern Oscillation. This criterion applies with additional force when either there is no

[6] The deterministic nature of the models make the rolling origin requirement more relevant because of the effects of the initial conditions.
[7] see http://en.wikipedia.org/wiki/Stylized_fact

suitable accuracy record available or the model is meant to apply in circumstances outside the range over which it was built, both of which obtain here. A potential problem arises from the sheer scale of model outputs which inevitably reveal some (possibly temporary) discrepancies between the model outputs and observed behaviour.

2.1.4 Black-Box and White-Box Validation

Because the AOGCMs are intended for use beyond the range of some of their input variables (e.g. most critically emissions) and their expected outputs (e.g. temperature) other validation criteria beyond comparative forecast accuracy come into play. This is relevant in order to understand the input – output relationships between the variables seen as primary causal inputs (in particular, emissions as they affect the output, temperature). Pidd (2003) remarks "[C]onfidence in models comes from their physical basis" and black-box validation based on input-output analysis should be supported by white-box (or open-box) validation. The aim is to demonstrate the observational correspondence with various sub-models, theoretically justified by science-based flow models as shown in the system in Figure 1 (e.g. emissions and atmospheric CO_2).,

The models have in part been designed to operate outside the domain of inputs from which they have been operationally constructed (i.e. the initial conditions and corresponding observations on temperature cannot include emissions at double the current level). Thus, it is important that the models demonstrate robust and plausible dynamic responses to inputs outside the observed range. The 'Climateprediction.net' experiment has been used to deliver some evidence on both model and initial condition sensitivity to a doubling of CO_2 (Stainforth et al., 2005), the results showing extremes of response (even including cooling). The experiment was also used to examine joint parameter sensitivity compared to the effects of single parameter tests. The former are needed, as here, because the effects may be more than the sum of the individual sensitivities.

2.1.5 Process Validation

The scientific community has developed its own procedures for assessing the validity of the models they develop. It depends primarily on peer review and replicability through open access to the proposed models, the data on which they are based and the models' outputs. At the heart of the processes is the concept of falsifiability (Popper 1959, 2002, but see Oreskes et al., 1994 and Kleindorfer et al. 1998 for a more focussed discussion relevant to GCMs) through critical predictive tests and replicability. Openness in making both data and models available is at the heart of both criteria. However, the peer review process acts as a gateway to gaining access to the high-performance computers required for replication and experimentation. In addition, a dominant consensus on how climate phenomena should be modelled can limit the range of models regarded as worthy of development (Shackley et al., 1998)[8].

The existence of a scientific consensus is in itself no guarantee of validity (Lakatos, 1970) and can in fact impede progress as ad hoc auxiliary hypotheses are added to shore up the dominant theory against empirical evidence. How monolithic is the GCM community of modellers? This issue has been addressed in an exchange between Henderson-Sellers and McGuffie (1999) and Shackley et al. (1999) with the latter arguing that despite different styles of modelling the predominant approach is 'deterministic reductionist', that is to say the models as described here (rather than, for example, aggregate statistical). Thus, no established critique of this modelling approach has developed either from inside or outside the global climate-modelling community.

2.1.6 Climate Scientists' Viewpoints on Model Validation

The IPCC Report contains the most authoritative views by climate scientists on model validation, often with a detailed discussion of the issues raised above (Le Treut, et al., 2007). The authors recognize all these elements of model validation, summarising both the process elements and the predictive requirement for model validation in Chapter 1 as follows, "Can the statement under consideration, in principle, be proven false? Has it been rigorously tested? Did it appear in the peer-reviewed literature? Did it build in the existing research record where appropriate?" and the results of failure are "less credence should be given to the assertion until it is tested and independently verified". The perspective the authors adopt is one where cumulative evidence of all types discussed above is collected to discriminate between one model (or explanation) and another but accepting a pluralistic (multi-model) perspective as reasonable practice (Parker, 2006). It is wholly compatible with the long-established but unacknowledged literature on the implications of the philosophical foundations of simulation-

[8] At the time of writing in February 2010, there is a major controversy arising from the theft of emails from the University of East Anglia, UK, as to the openness of the climate modelling community with regard to data and also the possible hijacking of the peer review process when there is a dominant paradigm (as here).

model validation for model-building practice (see Kleindorfer et al., 1998 for a survey and update).

Perhaps unfortunately, chapter 8 of the IPCC report, “Climate models and their evaluation”, (Randall et al., 2007) has not followed such a clear epistemological position. In particular, its view of falsifiability based on the analysis of in-sample evidence is overly limited in the criteria it lays down for its assessment of the AOGCM models “against past and present climate: section 8.1.2.3”. In fact the report backs away from model comparison and criticism, arguing the “differences between models and observations should be considered insignificant if they are within [unpredictable internal variability and uncertainties in the observations]”. Knutti (2008), for example, claims that “[A]ll AOGCMs… reproduce the observed surface warming rather well” despite robustness tests of parameters and initial conditions showing a wide range of simulated forecasts.

Chapter 8 also offers quite detailed evidence on various of the sub-models as part of open-box validation. There is little discussion of the input-output relationships. A related issue, although not in itself delivering direct evidence of the validity of the IPCC forecasts, is the use of ‘Earth System Models of Intermediate Complexity: EMICS’ which model aspects of the climate system by making simplifying assumptions about some of its elements, e.g. zonal averaging over geographical areas. Based on a keyword search of the 8 EMIC models listed in Chapter 10, *Global climate projections* (Meehl, et al., 2007) using the model name + forecast + valid in Google Scholar, the models have apparently not been used to provide forecast comparisons.

The discussion on model validation has moved on somewhat since the IPCC report of 2007, with greater emphasis on conformity of models with observation. Quite recently research programs have been developed by climate modelers to compare models (e.g. the Program for Climate Model Diagnosis and Intercomparison, Phillips et al, 2006) and to examine forecasting accuracy (Smith et al, 2007; Haines et al, 2009). Results from comparing models have shown that a combination of forecasts from different models is more effective than a single model (see for example, Hagerdorn et al., 2005) and that the improvement from adopting a multi-model approach is larger than that derived from using an ensemble of initial conditions in a single model. The individual model errors can potentially inform as to where improvements might lie.

In summary, the evidence provided in the IPCC report on the validity of the various AOGCMs, supplemented by much research work mostly from scientists within that community rests primarily on the physical science of the sub-models rather than their predictive abilities. The models also capture the stylised facts of climate such as the El Niño-Southern Oscillation. While the IPCC authors note that there is considerable agreement in the outputs of the various models, the forecasts do differ quite substantially and the combined model forecasts apparently conform better to recent data. The omissions of Chapter 10 of the IPCC and most of the subsequent research lies in the lack of evidence that the models produce good forecasts. There is ample testimony in the forecasting literature of the difficulties of forecasting beyond the range of data on which a model is constructed. This is only tempered slightly by the recognition that the physical sub-models are supposedly robust over the increasing CO_2 emissions input. In fact, climate modellers have raised 'completeness' in model building above all other criteria. It is not a criterion that earlier simulation modellers have ever regarded as dominant (Kleindorfer et al., 1998), rather it is often a diversion that detracts from understanding and forecast accuracy.

2.1.7 Outstanding Model Validation Issues

Despite the siren voices that urge us to reject the proposition that models can usefully be used in long-term forecasting (Oreskes, 2003), both the climate modelling community and forecasters share the belief that model based forecasts, whether conditional or unconditional, may provide information valuable in policy and decision making.

As forecasters examining the evidence, we have been struck by the vigour that various stylized facts and the 'white-box' analysis of sub-models are debated. An interesting example is that of tropospheric temperatures - Douglas et al. (2007) highlight a major discrepancy with model predictions followed by Allen and Sherwood (2008) critiquing their conclusions with web discussion contesting the proposed resolution. Where the debate has been most lacking is in the emphasis and evidence on forecast accuracy and forecast errors of the various models, although the discussion and initiatives described by Meehl et al. (2009) offer a welcome development. The AOGCMs themselves produce different forecast, both aggregate and regional, for key policy-relevant variables. The evaluation of these forecasts and their error distributions is potentially important for influencing the policy discussions. Issues such as the relative importance of mitigation strategies versus control (of emissions) depend on the validity of alternative models and the accuracy of their corresponding forecasts. Without a successful demonstration of the forecasting accuracy of the GCMs it is surely hard to argue that policy recommendations from such models should be acted upon. The study of forecasting accuracy of the models is a necessary (though not sufficient) condition for such models to guide policy and in the next section we will consider how climate forecasts from AOGCMs can be appraised with a view to improving their accuracy, focusing on the policy relevant variable of temperature.

3. Empirical Evidence on Forecast Accuracy

With so many requirements for model validation and possibilities of confusion, why, we might wonder, has the climate-change movement gained so much ground, despite considerable entrenched and powerful opposition? From a long-term perspective, there has been considerable variability in the Earth's climate. Examination of the ice-core record of Antarctic temperatures suggests a range of ±5ºC over the past 400,000 years as can be seen in Figure 2. However, changes of the magnitude of more than ±2ºC in a century have only be observed once, five centuries ago. Is the observed (but recent) upward trend shown in Figure 3, nothing more than an example of the natural variability long observed (Green and Armstrong, 2009) or is the projected temperature change in the IPCC report exceptional?

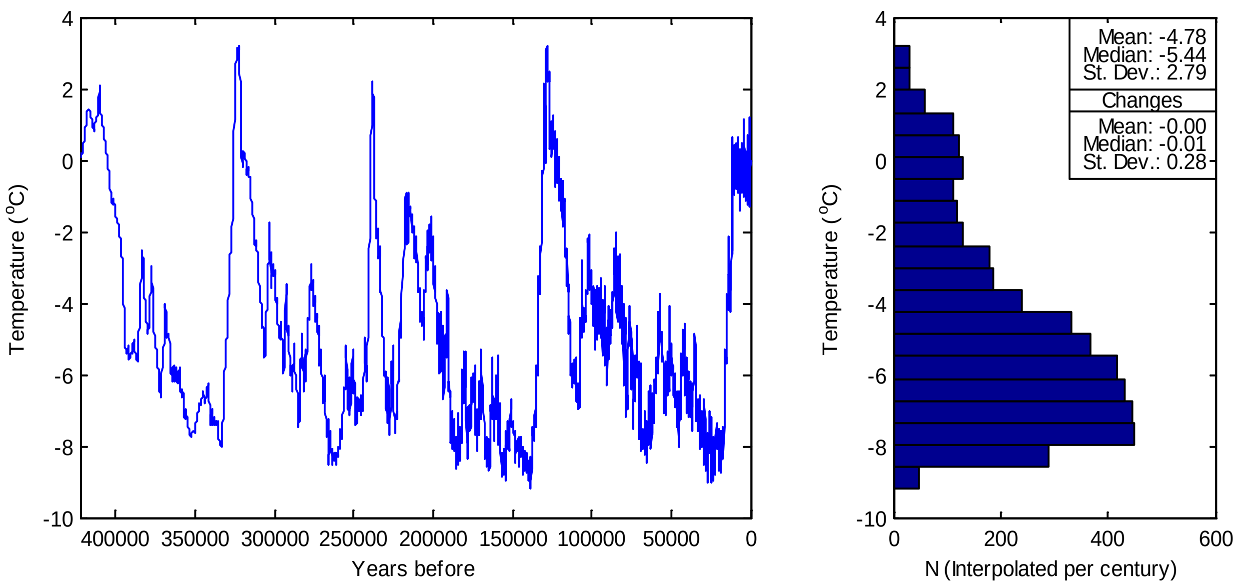

Figure 2 Vostok ice core temperature estimate plot and histogram. Histogram's values where missing are interpolated at century intervals to provide time equidistant estimations. (Data & Paper Reference Petit et al., (1999))

For the annual data needed for decadal modelling, there are many time series data of aggregate world temperature but it is only since 1850 that there have been data regularly collected that are broadly reliable – HadCRUT3v is the latest version of a well-established and analysed series, used to appraise AOGCMs. More recently, NASA[9] produced alternative estimates which have a correlation of 0.984 with the HadCRUT3v dataset. Since our focus here is on decadal climate change (up to 20 years), a long data series is needed and we have therefore used the Hadley Centre data in model building. In making this choice, we pass over the question of whether this series offers an accurate and unbiased estimate of global temperature. While the resolution of this uncertainty is of primary importance to establish the magnitude and direction of temperature change, it does not affect our methodological arguments directly. Figure 3 shows a graph of the HadCRUT3v data together with a 10-year moving average.

The features of the global temperature time series (the stylised facts) are of relative stability from 1850 through 1920; rapid increase until 1940 followed by a period of stability until 1970, after which there has been a consistent upward trend. From the longer-term data series such as the ice-core records, we can see that the bounds of recent movements (in Figure 3 ±0.6ºC) have often been broken. We conclude therefore that the temperature time series has seen persistent local trends with extremes that are both uncomfortably hot and cold (at least for humans). As we argued in section 2.1.2 on forecast validation, an important if not essential feature of a good explanatory model is its ability to explain such features of the data where other models fail. In particular, global climate models should produce better forecasts than alternative modelling approaches (in the sense that they are more accurate for a variety of errors measures)[10]. Over the time scales we are concerned with, therefore, any forecasting model should permit the possibility of a trend if it is to capture this particular feature of the data.

[9] http://data.giss.nasa.gov/gistemp/tabledata/GLB.Ts+dSSR.txt

[10] Perhaps some of the scepticism as to global warming is the failure of the IPCC to clearly demonstrate such success. There are of course a number of alternative hypotheses as to the underlying reasons for rejecting an apparent scientific consensus on global warming, starting with an unwillingness to listen to ‘bad news’.

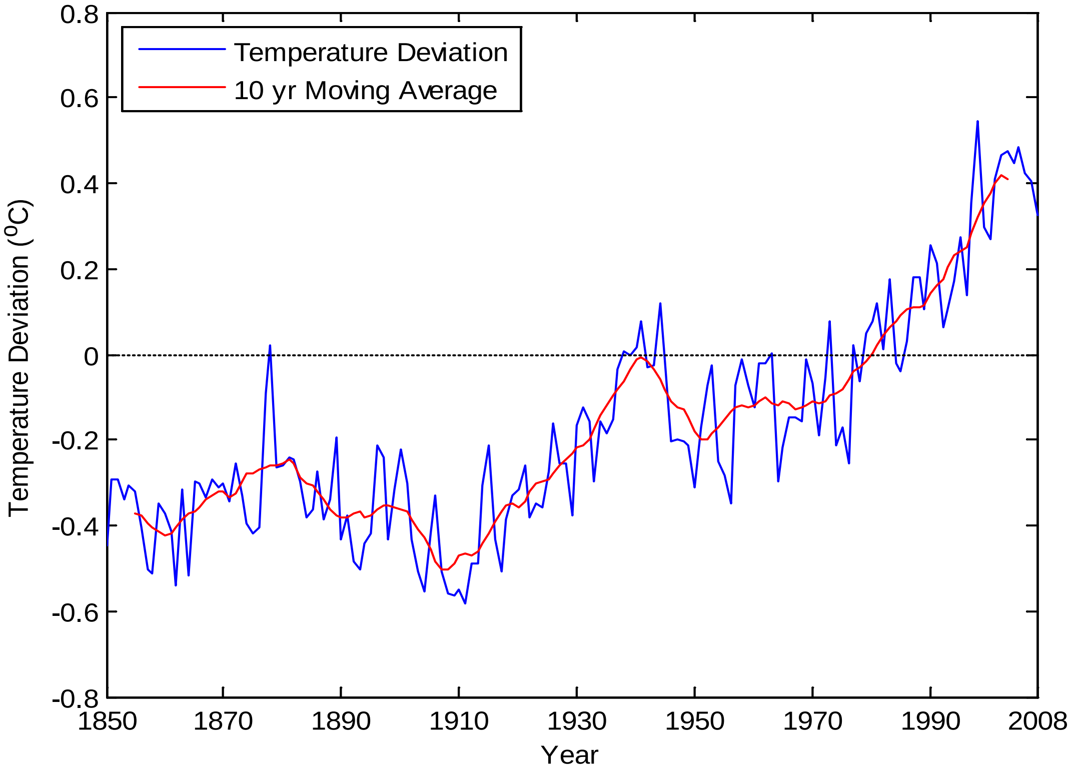

Figure 3 Temperature deviations °C and a ten year moving average (data taken from http://www.cru.uea.ac.uk/cru/info/warming/gtc2008.csv). NB. There has been some considerable controversy since at least some of the current data series suggest that the last few years have seen a decrease in global average temperature (see for example: http://tamino.wordpress.com/2008/04/16/perjury/#comments).

The evaluation of forecasts produced by GCMs requires a time series history, but this is not straightforward, since there is no established long historical record of forecasts that is definitive. However, we are able to benefit from Smith et al.'s (2007) work that provides us with a 25-year history of out-of-sample forecasts. The Smith study used a "newly developed Decadal Climate Prediction System (DePreSys), based on the Hadley Centre Coupled Model, version 3e (HadCM3)" specially designed to generate decadal predictions that would also take into account the initial conditions at the forecast origin. Only 1-10 year-ahead forecasts are currently available. Smith and his colleagues produced the forecasts as follows.

1. The model is run with pre-industrial levels of greenhouse cases as inputs until it reaches ‘steady climatological state’ – the control run. Most parameters (including constants) are theoretically or experimentally fixed. A number of parameters describe processes that are not fully specified and are chosen with reference to model behaviour. The initial conditions needed for the solution to the model are derived from an observed climatology but the effects of the choice die off over time.
2. An ensemble of (four) paths is generated using the natural variability observed in the control run (based on conditions taken 100 years apart to represent natural climate variability).
3. The model is now run from 1860 including observed greenhouse gases, changes in solar radiation and volcanic effects up to 1982Q1 to simulate the climate path.
4. The observed conditions for 4 consecutive days around the forecast origin are assimilated into the model to produce forecasts up to 10 years ahead, with forecasts based on observed forcings (with volcanic forcings only included when known ex post).

A fuller technical description is given in the on-line supporting material (Smith et al., 2007).

The essential difference between these forecasts and the standard simulation is that “atmospheric and ocean observations” on four consecutive days including the forecast origin, were used to produce the ten years ahead forecasts. When compared to forecasts produced by HadCM3 which did not take into account the observed state of the atmosphere and ocean the results (unsurprisingly) were substantially better.

The forecasts from the DePreSys model permit comparison with benchmark time series forecasts for the policy-relevant forecast horizon. The logic of this comparison is that it clarifies whether the GCM forecasts are compatible with the ‘stylised forecasting facts’ (of trend or no trend). If a trending univariate benchmark measured is more accurate ex ante accurate than the naive-no-change benchmark argued for by Green and Armstrong (2007) amongst others, this gives support to the notion of global warming. (Of course it tells us nothing about its causes or possible effective policy responses.)

The DePreSys forecasts are conditional forecasts based on various anthropomorphic variables, in particular $C0_2$ emissions. Using annual emissions from 1850 to 2006[11] (and an ARIMA(1,1,0) in logs to produce the forecast values of CO_2) we can construct multivariate models and carry out the same comparisons with the DePreSys forecasts and the univariate benchmarks. This gives us the potential to discriminate between the various effects embodied in the different benchmark models, thereby pointing the way to possible improvements in the Hadley GCM model. It also gives some information on whether CO_2 emissions can be said to Granger-cause Global temperature.

3.1. Evaluating Alternative Benchmarks

The results of the forecasting competitions (Fildes and Ord, 2002; Makridakis and Hibon, 2000) suggest various benchmark univariate forecasting methods. In addition, we consider both a univariate and a multivariate non-linear neural net. The data set used in model building is the annualised HadCrut3v and Total carbon emissions from fossil-fuels from 1850 to the forecast origin. We consider a number of forecast origins from 1938 to 2006. The estimation sample was extendedforward with each new forecast origin and the models were re-estimated. Forecast horizons were considered from 1 to 20, which are then separated into short-term and long-term forecasts.

The random walk offers the simplest benchmark model and for some types of data (e.g financial) it has proved hard to beat. In addition, Green and Armstrong (2007, 2009) have provided arguments for its use in climate forecasting although over the forecast horizons we are considering here (10-20 years) we do not regard them as strong. In addition we will try a number of benchmarks which have proved better than the naïve in the various competitions: simple exponential smoothing, Holt's linear trend and damped trend (Gardner, 2006). The latter two incorporate the key stylised fact of a changing trend. They have been estimated using standard built-in optimisation routines in MatLab®. Smoothing parameters and initial values were optimised using a MAE minimization of the estimation sample. We also consider simple linear autoregressive models with automatic order specification based on BIC optimisation[12]. These methods are all estimated on the time series of temperature anomaly changes.

[11] Global Fossil-Fuel CO_2 Emissions, Total carbon emissions from fossil-fuels (million metric tons of C), http://cdiac.ornl.gov/trends/emis/tre_glob.html

[12] A maximum lag of up to 25 years was used in specifying the AR models, similar to the univariate NNs.

In addition we have considered both a univariate and a multivariate neural network model (NN). Unlike the other models, they have the potential to capture non-linearities in the data. The univariate NN is modelled on the differenced data and the inputs are specified using backward dynamic regression[13], evaluating lag structures up to 25 years in the past. For the case of the multivariate NN a similar procedure is used to identify significant lags of the explanatory variable, considering lags up to 15 years in the past. No contemporaneous observations are used. We use a single hidden layer. To specify the number of hidden nodes in the layer there is no generally accepted methodology (Zhang, et al., 1998), therefore we perform a grid search from 1 to 30 hidden nodes. We identified 11 and 8 nodes to be adequate for the univariate and the multivariate NN respectively. Formally the model is

$$f(X, w) = \beta_0 + \sum_{h=1}^{H} \beta_h g\left(\gamma_{h0} + \sum_{i=1}^{I} \gamma_{hi} x_i \right)$$

where $g(x) \equiv \tanh(x) = \dfrac{2}{\left(1 + e^{-2x}\right) - 1}$ (Vogl, et al., 1988):

The hyperbolic tangent activation function in the hidden nodes is used to model nonlinearities in the time series. There is a single linear output that produces a t+1 forecast. Longer forecasting lead times are calculated iteratively. For the training of the NNs we split the in-sample data into training and validation subsets. The last 40 observations constitute the validation set and the remaining observations the training set. The NNs are trained using the Levenberg-Marquardt algorithm, minimising the 1-step-ahead in-sample mean square error. Each NN is randomly initialised 20 times to mitigate the problems that arise due to the stochastic nature of NNs' training. The final forecast is calculated as the median output of all these 20 different initialisations. Finally the NNs are retrained at each origin. We have used a black-box input-output approach for the multivariate neural nets using for inputs CO_2 annual emissions and lagged values of the temperature anomaly. Volcanic emissions have been excluded, ensuring the results are comparable to Smith et al.'s.

The final forecasting method considered is based on combining the forecasts from all the other methods, giving equal weight to each method.

[13] A regression model is fitted and the significant lags are used as inputs to the neural network (Kourentzes, N., & Crone, S.F. 2010. Input variable selection for forecasting with neural networks. *Lancaster University Management School*. Lancaster, UK: Lancaster University..

The primary forecast horizon is the 10- and 20-years-ahead temperature deviation with absolute error as the corresponding error measure. However, the compatibility between the shorter-term horizons (we will consider 1-4 years) and the longer horizons also offers evidence of model validity.

3.1.1 Short term forecasting results

Table 1 summarises the 1-4 year ahead mean (median) absolute errors from the various models: the random walk, simple exponential smoothing, Holt's linear trend, a damped trend model, the AR model, the univariate and multivariate NN models and the combination of forecasts, and for different hold-out samples. They are compared to Smith et al.'s forecasts where possible.

	MAE (MdAE) in forecasting 1-4 years ahead			
	Method	Hold-out sample period		
		1939-2007	1959-2007	1983-2005
Horizon 1-4	Naive	0.109 (0.094)	0.108 (0.094)	0.116 (0.100)
	Single ES	0.104 (0.103)	0.099 (0.092)	0.106 (0.101)
	Holt ES	0.122 (0.104)	0.104 (0.091)	**0.084** (0.082)
	Damped Trend ES	0.115 (0.101)	0.097 (0.085)	0.098 (0.089)
	AR	0.109 (0.093)	0.107 (0.093)	0.113 (0.097)
	NN-Univariate	0.104 (0.089)	0.096 (0.083)	0.094 **(0.080)**
	NN-Multivariate	0.101 **(0.084)**	0.097 **(0.079)**	0.098 (0.093)
	Combination	**0.099** (0.092)	**0.091** (0.089)	0.092 (0.091)
	Smith	-	-	0.163 (0.160)
No. of observations		66	46	20

Table 1 Mean and Median Absolute Error (MAE and MdAE) in Forecasting 1-4 years ahead Average Global Temperature Deviation using alternative univariate and multivariate forecasting methods, compared to Smith et al.'s GCM forecasts. The most accurate method(s) are shown in bold.

The short-term forecasts show high variability in performance of the various extrapolative models. Thus, the combined forecast performs well. The NNs performs well over the longer data base but the more consistent upward trend in the last 20 years allowed Holt's local linear trend model to beat them. The forecasts from DePreSys performed poorly.

3.1.2 Longer-term forecasts

Table 2 shows the results for similar comparisons for the 10-year-ahead forecasts and the average of the one-year through ten-year forecasts. Where comparison with Smith is possible, we see that while the GCM model performs well compared to the simple benchmark alternatives, the NN models and Holt's forecasts have similar or better performance. The neural networks and the combined forecasts performed overall the best when evaluated over long hold-out periods. Holt's model outperforms the rest for during 1983-2005 period when there is a significant trend in the data.

Table 3 provides the results for 11-20 and 20-years-ahead forecasts. While there are no forecasts for DePreSys the multivariate NN that considers CO_2 information performs consistently the best in long term forecasting for a sample of the last 30 years of the holdout sample. This effect becomes more apparent during the last decade, where the errors of the multivariate NN become substantially lower than all other models.

Method	MAE (MdAE) in forecasting 1-10 and 10 years ahead Hold-out sample period: Horizon 1-10			Horizon 10		
	1939-2007	1959-2007	1983-2007	1948-2007	1968-2007	1992-2007
Naive	0.130 (0.113)	0.129 (0.109)	0.148 (0.133)	0.152 (0.142)	0.155 (0.142)	0.202 (0.198)
Single ES	0.126 (0.116)	0.126 (0.111)	0.153 (0.144)	0.156 (0.130)	0.168 (0.160)	0.220 (0.242)
Holt ES	0.153 (0.129)	0.118 (0.109)	**0.091** (0.092)	0.184 (0.146)	0.136 (0.125)	**0.088** (0.084)
Damped Trend ES	0.136 (0.124)	0.122 (0.113)	0.138 (0.138)	0.158 (0.134)	0.161 (0.145)	0.195 (0.189)
AR	0.128 (0.110)	0.123 (0.107)	0.134 (0.122)	0.140 (0.122)	0.131 (0.119)	0.169 (0.156)
NN-Univariate	0.122 **(0.096)**	**0.102 (0.086)**	0.097 **(0.074)**	0.136 **(0.091)**	**0.106 (0.087)**	0.098 (0.079)
NN-Multivariate	0.124 (0.098)	0.115 (0.091)	0.100 (0.090)	0.154 (0.136)	0.131 (0.099)	**0.088** (**0.058**)
Combination	**0.118** (0.105)	0.105 (0.097)	0.112 (0.104)	**0.133** (0.113)	0.118 (0.110)	0.133 (0.131)
Smith	-	-	0.140 (0.132)	-	-	**0.088** (0.064)
No. of observations	60	40	16	60	40	16

Table 2 Mean and Median Absolute Error (MAE and MdAE) in Forecasting 10 years ahead Average Global Temperature Deviation using alternative univariate and multivariate forecasting methods, compared to Smith et al.'s GCM forecasts.

A graph of the ratio of time periods where the random walk (RW) is the winner is provided in figure 4 and shows a clear story. The ratio for each method is calculated in the following way; a vector is created that contains 1 if the random walk is better and 0 otherwise for each year. Then for each year to find the ratio, the sum of the vector up to that date is divided by the number of years of measurement, therefore a value of 1 would mean that the random walk always wins, 0 the opposite and any values in-between would show preference towards the respective model. For our calculations we use the 10-years-ahead forecasts. From around 1970 there is a clear preference for models that are able to model trend and the ratio turns against the random walk.

Method	MAE (MdAE) in forecasting 11-20 and 20 years ahead Hold-out sample period					
	Horizon 11-20			Horizon 20		
	1949-2007	1969-2007	1993-2007	1958-2007	1978-2007	2002-2007
Naive	0.182 (0.163)	0.215 (0.197)	0.312 (0.305)	0.202 (0.181)	0.273 (0.276)	0.386 (0.413)
Single ES	0.176 (0.152)	0.226 (0.214)	0.332 (0.350)	0.208 (0.182)	0.290 (0.310)	0.406 (0.404)
Holt ES	0.283 (0.263)	0.236 (0.233)	0.175 (0.165)	0.355 (0.301)	0.306 (0.284)	0.195 (0.251)
Damped Trend ES	0.192 (0.170)	0.226 (0.214)	0.330 (0.353)	0.230 (0.192)	0.287 (0.315)	0.402 (0.406)
AR	0.173 (0.148)	0.184 (0.164)	0.259 (0.262)	**0.178 (0.134)**	0.220 (0.207)	0.312 (0.344)
NN-Univariate	0.174 (0.138)	**0.144** (0.127)	0.160 (0.132)	0.200 (0.146)	0.175 (0.139)	0.203 (0.210)
NN-Multivariate	0.179 (**0.138)**	**0.144 (0.102)**	**0.105 (0.078)**	0.195 (0.149)	**0.131 (0.103)**	**0.125 (0.111)**
Combination	**0.166** (0.154)	0.171 (0.162)	0.226 (0.229)	0.194 (0.181)	0.212 (0.235)	0.267 (0.273)
Smith	-	-	-	-	-	-
No. of observations	50	30	6	50	30	6

Table 3 Mean and Median Absolute Error (MAE and MdAE) in forecasting 10 years ahead Average Global Temperature Deviation using alternative univariate and multivariate forecasting methods.

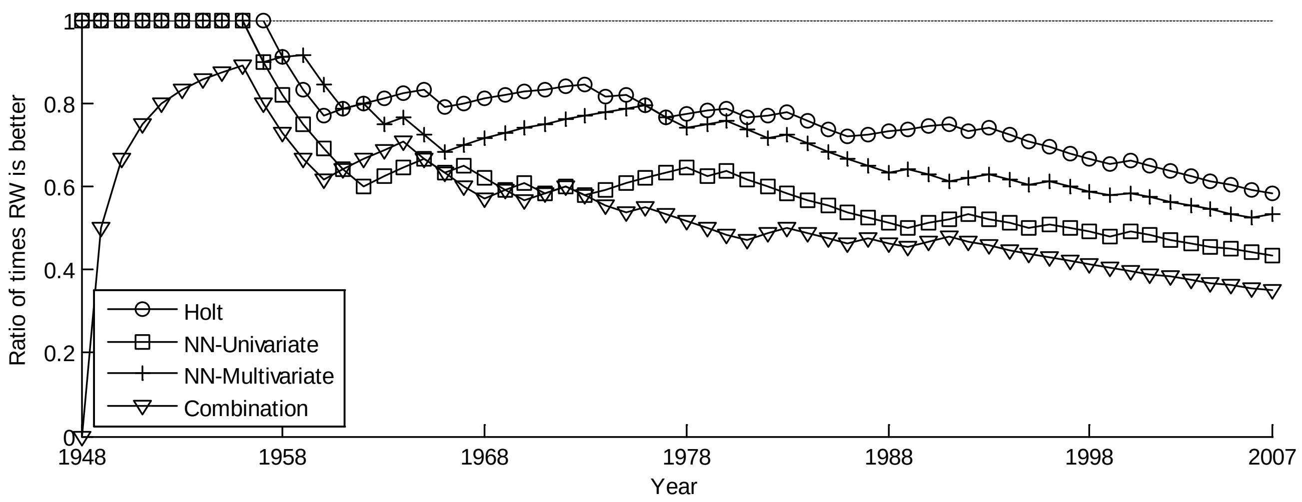

Figure 4 Cumulative ratio of proportion of periods that the random walk is the winner for 10 year ahead forecasts.

(If the ratio tends towards one the random walk is preferred and if it tends towards zero the opposite. Note that for the first 9 years all methods but the combined forecast are always worse than the random walk. From 1960s on the other forecasts methods start to outperform the random walk and the cumulative ratio averaged over the different forecasting origins drops.)

The unconditional forecasts for the 10-year and 20-year ahead world annual temperature deviation range between 0.1- 0.2°C per decade for the methods that are able to capture trends. This compares with the best estimate from the various climate change models of .2°C (approx) for the A2 emissions scenario. The forecasts for all models are provided in figure 5 and a summary in table 4. Note that the models that accurately predict the temperature, forecast increases in the temperature for the last two decades. In comparison with the A2 scenario[14] from the IPCC AR4 report, the NN-Multivariate model provides the same per year temperature increase forecast.

Method	Year		Change per decade (°C)		Trend estimation per decade (°C)
	2017 (t+10)	2027 (t+20)	2017 (t+10)	2027 (t+20)	
Naive	0.398	0.398	0.000	0.000	0.000
Single ES	0.421	0.421	0.023	0.000	0.000
Holt ES	0.702	0.913	0.304	0.211	0.211
Damped Trend ES	0.615	0.709	0.217	0.094	0.118
AR	0.451	0.505	0.053	0.053	0.053
NN-Univariate	0.357	0.050	-0.041	-0.307	-0.042
NN-Multivariate	0.559	0.748	0.161	0.189	0.180
Combination	0.501	0.535	0.103	0.034	0.074
IPCC AR4 Scenario A2					**0.180**
2007 Observed Temperature deviation					0.398

Table 4 Unconditional forecasts for 10 and 20 years ahead world annual temperature deviation.

[14] This scenario assumes regionally oriented economic development with no environmentally friendly policies implemented, simulating the current conditions.

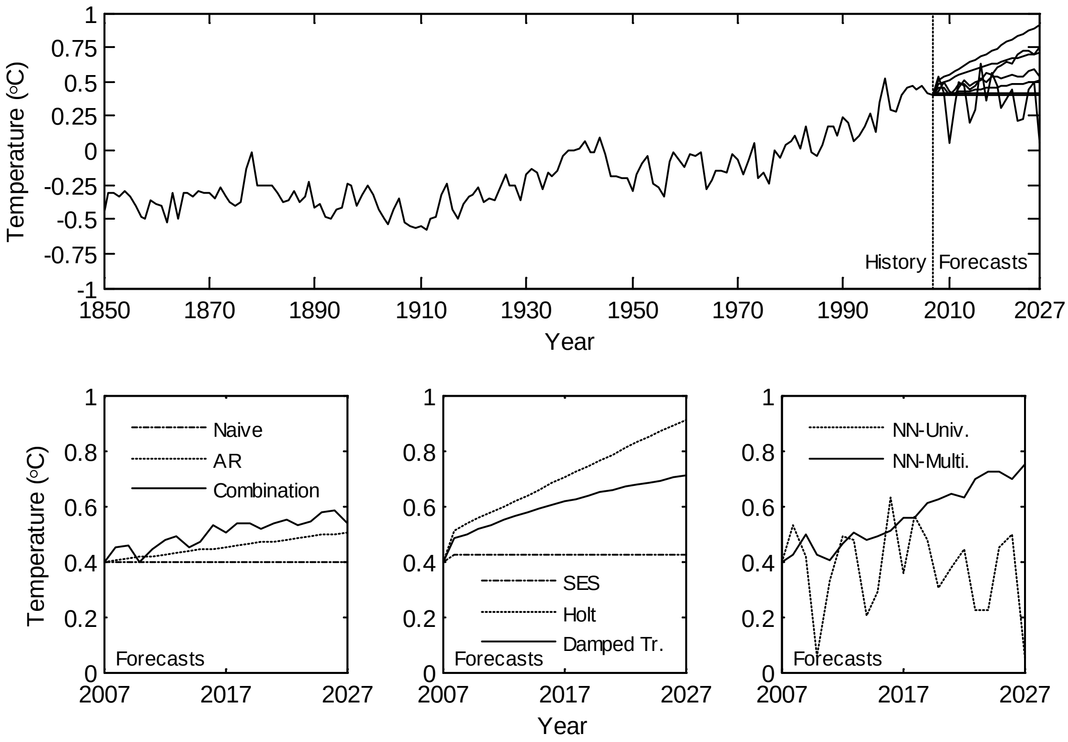

Figure 5 20-years-ahead world annual temperature deviation forecasts for all methods.

However, the above analysis does not say anything about the causes of the trend (or even anything much about global warming). It does however show the trend to continue over the next ten or twenty years. It is also quite persistent in that the full data history shows there are relatively few rapid reversals of trend. By plotting the changes in the trend component of the 10-years-ahead Holt's forecasts, in figure 6, which provides a smooth estimation of the historical trend, we can observe that the trend estimate remains relatively low and there are very few years with negative trend.

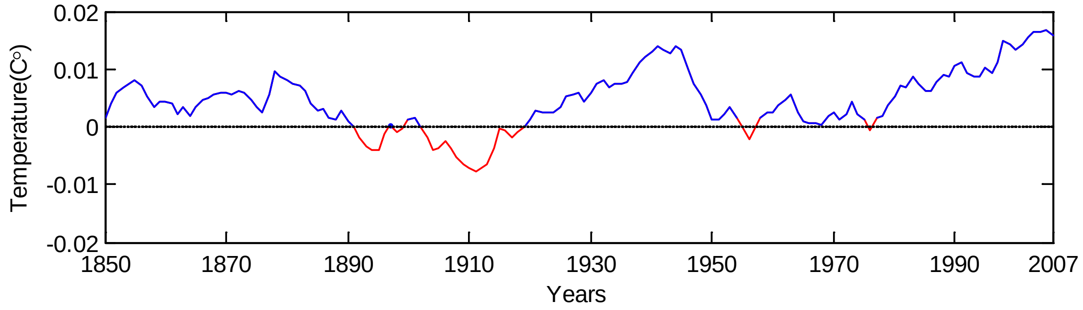

Figure 6 Trend component estimation of the temperature deviation from the 10-year-ahead in-sample Holt forecast

3.2. Encompassing tests

A forecast encompassing test of the DePreSys forecasts compared to the other forecasting methods allows us to test whether the various benchmarks we considered in the previous section add additional information and which are the most valuable.

Formally, there are a number of models that can be used as the basis of encompassing tests (Fang, 2003). We examine three variants

$$Temp_t = \alpha ForMeth1_{t-h}(h) + (1-\alpha) ForMeth2_{t-h}(h) + e_t \qquad (M1)$$

$$Temp_t = \alpha_0 + \alpha_1 ForMeth1_{t-h}(h) + \alpha_2 ForMeth2_{t-h}(h) + e_t \qquad (M2)$$

$$Temp_t - Temp_{t-h} = \alpha_0 + \alpha_1 (ForMeth1_{t-h}(h) - Temp_{t-h}) + \alpha_2 (ForMeth2_{t-h}(h) - Temp_{t-h}) + e_t \quad (M3)$$

Where *Temp* is the actual temperature and $ForMethi_{t-h}$ (*h*) is the h-step-ahead forecast produced in period t-h using method i. Equation (1) is the standard combining approach which can also be used to test for encompassing through the test for α=0 (or=1). Equation (2) permits the possibility of bias and is due to Granger and Ramanathan (1984). The third recognizes the possibility of non-stationary data (Fang, 2003). This can be examined in an unconstrained or constrained form, where α_1 and α_2 must add up to 1, as in equation (1). Here we examine the constrained case.

In table 5 we present the 10-year- and 20-year-ahead forecasts. Our focus is on establishing which, if any, methods encompass the others. In part this question can be answered by considering the theoretical basis of the models. We will therefore only consider pairs of methods that have distinct characteristics. The pairs we will therefore consider are taken from the following: AR, Exponential Smoothing, univariate Neural Network and multivariate Neural Network. Holt's Linear trend model has been chosen from the exponential smoothing class as having the lowest correlations with the other methods and support for this was found through a varimax factor analysis with three factors.

Type	Methods	Horizon 10			Horizon 20		
		1948-2007	1968-2007	1992-2007	1958-2007	1978-2007	2002-2007
Model 1	AR & Holt	0.172 (A)	0.143 (AB)	0.107 (B)	0.225 (A)	0.261 (A)	0.238 (-)
	AR & NN univ.	0.169 (A)	0.129 (B)	0.105 (B)	0.224 (A)	0.232 (B)	0.159 (B)
	AR & NN multi.	**0.167 (AB)**	0.144 (AB)	0.123 (A)	**0.199 (AB)**	0.160 (AB)	0.276 (-)
	Holt & NN univ.	0.184 (B)	0.120 (AB)	0.101 (-)	0.272 (B)	0.236 (B)	**0.155 (B)**
	Holt & NN multi.	0.174 (AB)	**0.116 (AB)**	**0.090 (AB)**	0.231 (AB)	**0.141 (AB)**	0.212 (A)
	NN univ. & NN multi.	0.176 (AB)	0.125 (AB)	0.111 (A)	0.231 (AB)	0.154 (AB)	0.156 (A)
Model 2	AR & Holt	**0.168 (A)**	**0.092 (AB)**	0.094 (B)	0.205 (A)	**0.118 (AB)**	0.133 (-)
	AR & NN univ.	0.169 (A)	0.117 (AB)	0.104 (-)	0.205 (A)	0.143 (AB)	0.168 (-)
	AR & NN multi.	**0.168 (A)**	0.132 (A)	0.115 (A)	**0.200 (A)**	0.151 (A)	**0.120 (-)**
	Holt & NN univ.	0.185 (B)	0.095 (AB)	0.096 (A)	0.274 (B)	0.135 (AB)	0.162 (-)
	Holt & NN multi.	0.173 (AB)	0.103 (AB)	**0.092 (A)**	0.224 (B)	0.122 (AB)	0.122 (-)
	NN univ. & NN multi.	0.171 (AB)	0.124 (A)	0.115 (A)	0.222 (B)	0.151 (AB)	0.136 (-)
Model 3	AR & Holt	**0.168 (A)**	**0.092 (AB)**	0.094 (B)	0.205 (A)	**0.118 (AB)**	0.133 (-)
	AR & NN univ.	0.169 (A)	0.117 (AB)	0.104 (-)	0.205 (A)	0.143 (AB)	0.168 (-)
	AR & NN multi.	**0.168 (A)**	0.132 (A)	0.115 (A)	**0.200 (A)**	0.151 (A)	**0.120 (-)**
	Holt & NN univ.	0.185 (B)	0.095 (AB)	0.096 (A)	0.274 (B)	0.135 (AB)	0.162 (-)
	Holt & NN multi.	0.173 (AB)	0.103 (AB)	**0.092 (A)**	0.224 (B)	0.122 (AB)	0.122 (-)
	NN univ. & NN multi.	0.171 (AB)	0.124 (A)	0.115 (A)	0.222 (B)	0.151 (AB)	0.136 (-)
Number of observations		60	40	16	50	30	6

Table 5 Forecast encompassing tests of pairs of time series models based on Models M1-M3, 10 and 20 years ahead. Standard errors are reported. In parenthesis the significant forecasting methods are noted, A being the first, B the second and AB both under a 5% significance level.

Considering the results for M1 in both 10- and 20-years-ahead forecasts there is a consistent picture that the combination of neural networks and linear models (AR and Holt) provides the lowest standard error, implying that there are important nonlinearities in the data. Under M2 and M3 the picture is more complicated. Again the combination of neural networks and linear models provides useful synergies, however the combination of AR and Holt methods performs very well, especially for the 10-years-ahead forecasts. On longer horizons the importance of multivariate NNs is more apparent, providing some evidence that the effects of CO_2 become more prominent in longer horizons.

Looking ten years ahead we have some limited evidence of good performance from the DePreSys GCM forecasts. We consider a different model here, examining whether additional accuracy can be achieved through the additional information available from the statistical models. The proposed model is:

$$Temp_t = a_0 + \sum_{i=1}^{k} a_i ForMeth_{i,t-10}(10) + \lambda DePreSys_{t-10}(10) + e_t \qquad (M4)$$

Since the error is stationary, there is no reason to consider differences as in equation (*M3*). Essentially, a significant coefficient (to $ForMeth_i$) suggests that the GCM is failing to capture the key characteristic embodied in that particular forecasting method. The combination of forecasts can be done for 1 to k different methods. A significant constant term suggests a consistent bias. A significant coefficient of the forecasting method implies that there is additional information that is not captured by the GCM, HadCM3. If we take $\lambda=1$, this in effect poses the question as to whether the error made by the GCM can be explained (and improved upon) by other time series forecasting methods.

We present the results for combination of one method with DePreSys in table 6. All combinations demonstrate improvements over the individual forecasts of DePreSys, which have a standard error of 0.103. However, only Holt linear trend exponential smoothing forecasts seem to have a significant impact on improving the accuracy, implying that in the limited period that DePreSys forecasts were available the upward trend in temperature is not captured adequately. On the other hand, the nonlinearities modelled by the equally accurate NN models do not provide significant additional new information on the 10 years ahead forecast for that period, though the standard error of the combined forecast is improved. If the level and trend component of Holt's forecasts are considered separately then trend exhibits a significant coefficient of +1.128, resulting in a standard error of 0.087, marginally worse than Holt, further strengthening the argument that the DePreSys forecasts do not capture adequately the trend exhibited in the data. The level component is marginally insignificant with a coefficient of 0.564, resulting in a reduction of the standard error to 0.093.

Method	Constant	Method Coefficient	Standard Error
Naive	+0.022 (0.591)	+0.318 (0.206)	0.099
Single ES	+0.002 (0.959)	+0.581 (0.139)	0.096
Holt ES	**-0.088 (0.154)**	**+0.561 (0.014)**	**0.084**
Damped Trend ES	+0.033 (0.453)	+0.243 (0.357)	0.101
AR	-0.035 (0.597)	+0.367 (0.116)	0.095
NN-Univariate	-0.042 (0.582)	+0.338 (0.151)	0.097
NN-Multivariate	+0.013 (0.787)	+0.298 (0.207)	0.099
Combination	-0.029 (0.625)	+0.467 (0.098)	0.094
Smith	-	-	0.103

Table 6 Forecast error models of the DePrSys 10 year ahead Forecasts (1992-2007). Numbers in brackets are p-values.

To obtain the results for combinations of two or more methods the model is constrained so that the coefficients are positive. The findings are less interesting, since the Holt forecasts dominate the rest, forcing their contribution to zero or very close to zero. The unconstrained model does not permit easy interpretation merely pointing to the collinearity between the various forecasts.

3.3. Localised temperature forecasts

One important use of highly disaggregated GCMs is to produce local forecasts of temperature, rainfall, extreme events, etc. These are used by many agencies, both in government and commercially, to examine the effects of predicted climate change at a local level (see e.g., http://precis.metoffice.com/). In terms of forecast validation, they provide a further test-bed for understanding the strengths and deficiencies of the GCMs. Koutsoyiannis et al. (2008) have explored this issue by evaluating various GCMs used in both the third and fourth IPCC assessment reports. In brief, rainfall and temperature were measured from 8 localities from around the world. Six GCMs were then used to provide estimates of these quantities and the results compared on an annual basis using a variety of measures including comparisons of various summary statistics (mean, autocorrelation, etc.) and error statistics including the correlation between the observed and predicted values of rainfall and temperature and the coefficient of efficiency[15]. The simulations from the GCMs are not forecasts in the same way as Smith et al's carefully produced results, because they are not reinitialised at the forecast

[15] The coefficient of efficiency is used in hydrology and relates to R^2.
It is defined as $= 1 - \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$

origin. Such simulations are often interpreted in much the same way as forecasts, generating arguments and policy proposals that presume the simulated values have the same validity (or lack thereof) as forecasts. Koutsoyiannis et al's (2008) evaluate the GCM simulations at seasonal, annual and 30-year (climatic) time scales, measured through a 30-year moving average of the annual temperature. While the models capture seasonal variation, the results for the two longer horizons are uniformly negative. We have carried out some limited calculations to extend their results using a forecasting framework and standard error measures, which are less prone to misinterpretation. Here we compare our one and ten-year-ahead forecasts from our time series benchmarks with the GCM forecasts. The aim is to compare the 'stylised facts' in different localities with the observations and following on from our aggregate analysis to see if our time series forecasts add information to the local GCM based forecasts.

The simulations were run for six regions (Albany, Athens, Colfax, Khartoum, Manaus, Matsumoto) as in Koutsoyiannis et al. who provided us with the regional data and the GCM simulations. We use the Median Relative Absolute Error (MdRAE), which compares the accuracy of a method as a ratio to the Random Walk's accuracy for the same period and is calculated as:

$$MdRAE_{i,h} = Md\left(\frac{\left|Actuals_t - ForMeth_{i,t}\right|}{\left|Actuals_t - Actuals_{t-h}\right|}\right)$$

The closer an error is to zero the more accurate that method is and if it is equal to one then the method is only as good as the random walk. The reason we consider the MdRAE is that we present the results on an aggregate level across all six regions. Here we consider a combination of the GCM forecasts provided in Koutsoyiannis et al. The combination is done through averaging between the different predictions. The results were aggregated to give overall accuracy, allowing us to present the results for all regions in a summarised format; summary statistics are presented in table 7. The GCM models performed significantly worse than the random walk. However the benchmark forecasting methods used in this study exhibited similar or better performance to the naive. This implies that the current GCM models are ill-suited for regional decadal predictions, even though they are used as inputs to policy making.

Method	MdRAE	
	t+1	t+10
Naive	1.00	1.00
Single ES	0.87	**0.90**
Holt ES	0.99	1.14
Damped Trend ES	0.95	0.98
AR	0.95	0.98
NN-Univariate	1.02	1.02
NN-Multivariate	1.03	1.76
Combination	**0.85**	0.93
GCM	3.79	3.01

Table 7 MdRAE for 1- and 10-step-ahead temperature regional forecasts. Results are aggregate errors across all six regions[16]. (Data downloaded from http://climexp.knmi.nl/. Details from the authors)

Using the regional data we can also compare the relative forecasting performance of the methods to the random walk in a global and regional scale. This is done in table 8, where forecasting accuracy in terms of MdRAE for several horizons is provided both for the regional and global forecasts. It is apparent that most of the methods (with the exception of Single ES and Damped Trend ES) can capture and model additional structure in comparison with the naive for the global time series, resulting in significant improvements in accuracy in comparison with the regional forecasts. In contrast, for regional time series the gains of the methods over the random walk are marginal and in most cases they are unable to capture additional information that would result in significant accuracy improvements. Therefore, for horizons up to 20 years ahead, which are tested here, there is evidence that the no-change scenarios, which are expressed through the random walk, do not adequately describe the observations and there is additional structure, appearing as local trends and nonlinearities, which must be considered in forecasting the global temperature deviation.

[16] For each region data of different length are available, leading to different evaluation periods. For Albany the evaluation period is 45 year long (allowing for 45 t+1 and 36 t+10 forecasts). Similarly the evaluation period for Athens is 89, for Colfax 75, for Khartoum 46, for Manaus 37 and for Matsumoto 49 years long. The accuracy over each forecast horizon for each time series is calculated first and then aggregated over all regions.

	Method	t+1-t+4 1983-2005	t+1-t+10 1983-2007	t+10 1992-2007	t+20 2002-2007	t+11-t+20 1993-2007
Regional	Naive	1.000	1.000	1.000	1.000	1.000
	Single ES	0.812	0.850	0.992	**0.866**	0.948
	Holt ES	0.878	0.864	0.998	1.148	1.213
	Damped Trend ES	0.819	0.860	0.989	0.916	0.996
	AR	0.930	0.939	0.985	1.214	1.069
	NN-Univariate	0.860	0.867	1.035	1.397	1.349
	NN-Multivariate	0.842	0.817	0.946	1.385	1.272
	Combination	**0.802**	**0.813**	**0.913**	1.036	**0.882**
Global	Naive	1.000	1.000	1.000	1.000	1.000
	Single ES	0.833	0.995	1.064	1.016	1.012
	Holt ES	**0.699**	**0.601**	**0.362**	0.572	0.567
	Damped Trend ES	0.787	0.880	0.977	1.004	1.010
	AR	0.956	0.898	0.815	0.832	0.838
	NN-Univariate	0.793	0.632	0.505	0.439	0.386
	NN-Multivariate	0.806	0.654	0.402	**0.264**	**0.283**
	Combination	0.790	0.739	0.647	0.667	0.710
	Smith	1.591	0.904	0.406	-	-

Table 8 MdRAE accuracy for regional and global forecasts. Most accurate method for each horizon is marked in boldface.

4. Discussion and Conclusions

Decadal prediction is important from both the perspective of climate-model validation and for assessing the impact of the forecasts and corresponding forecast errors on policy. The results presented here show that current decadal forecasting methods using a GCM, whilst providing better predictions than those available through the regular simulations of GCMs (and the IPCC), have limitations. As yet there are only a limited number of 10-year-ahead forecasts, derived from a single GCM (HadCM3 using DePreSys). But based on these forecasts we have shown that overall forecast accuracy from the DePreSys could have been improved on. More importantly, through the combining and encompassing analysis, various model weaknesses were identified. In particular, the value of adding Holt's model to the DePreSys forecasts proved of some value (dropping the standard error by 18%). By decomposing Holt's model forecasts into its structural components of level and trend, we were able to demonstrate that both components add value to the DePreSys forecasts, that is, the re-initialisation that takes place at the forecast origin is inadequate. But the failure to capture the local linear trend is perhaps more surprising. Other forecasting methods, in particular neural nets add nothing to the GCM forecasts. In essence, this suggests that the GCM is capturing the non-linearities in the input-output response to emissions but that it fails to adequately capture the local trend. The decadal forecasting exercise is apparently over-reactive to the forecast origin with a smoothed value of the current system state from the exponential smoothing model providing more adequate forecasts.

The aim of this paper has been to discuss the claims made for the validity of GCMs as a basis for medium-term decadal forecasting and, in particular, to examine the contribution a forecasting research perspective could bring to the debate. As our analysis has shown, at least with regard to the DePrSys model, it provides 10-year-ahead forecasts that in aggregate could be improved by adding in statistical time series forecasts. At a more localised level, using simulations from a range of IPCC models that have not been data-assimilated at the forecast origin and are therefore less likely to provide accurate decadal predictions, we found very low levels of accuracy (as did Koutsoyiannis et al., 2008).

What do these comparative forecast failures imply for model validation? Within the climate modelling community there is a consensus surrounding the concept of a valid model and it is generally accepted that there can be no conclusive test of a model's validity. Instead various aspects of a model are evaluated and the results add support (or not) to that model. To overcome the realisation that all of the models used in the IPCC forecasts have weaknesses, a combined (ensemble) forecast is produced. However, comparative forecasting accuracy has not been given much prominence in the debate, despite its importance both for model validation and for policy (Green and Armstrong, 2007; Green et al., 2009). Our analysis has identified structural weaknesses in the model(s) that should point the way for climate researchers to modify either their model structure and parameterisation, or if the focus of the modelling exercise is decadal forecasting, the initialisation and data assimilation steps. We cannot over-emphasize the importance of the initiative described by Meehl et al. (2009), firmly rooted as it is in the observed state of the system at the forecast origin. This new development aims to provide accurate forecasts over a horizon of 10-30 years, a forecast horizon relevant for policy. In carrying out the analysis reported here we have achieved improvements of forecasting accuracy of some 35% for up to 10-year forecasts. Such improvements have major policy implications and consequential cost savings.

Extending the horizon of decadal forecasting using a GCM to 20 years is practical, although the computer requirements are extensive. We have also carried out some limited analysis of 20-year forecasts without, therefore, the benefit of any corresponding forecasts from a GCM.

While for the 10-year forecasts the signal is potentially lost in the noise, any trend caused by emissions or other factors (see for example Pielke et al., 2009) should be observed in the forecast accuracy results. In the 20-year-ahead forecasts the multivariate neural net was shown to have improved performance over its univariate alternatives. Interpreted as a Granger-causality test, the results unequivocally support the importance of emissions as a causal driver of temperature backed as it is by both scientific theoretic arguments and observed improvements in predictive accuracy. But there is no comfort for those who reject the whole notion of global warming– the forecasts still remain inexorably upward with forecasts that are comparable to those produced by the models used by the IPCC. The long-term climate sensitivity to a doubling of CO_2 from its pre-industrial base is not derivable from the multivariate neural net which is essentially a short-term forecasting model. A current review of estimates arrives at a value of around 2.8 with a 95% confidence interval of 1.5 to 6.2 (Royer, et al., 2007), compatible with figures from the IPCC models.

The observed short-term warming over recent decades has led most climate change sceptics to shift the terms of the political argument to the question of the climate's sensitivity to CO_2 emissions. Here we find a conflict between two aspects of model validation – the criterion of providing more accurate forecasts than those from competing models, and the other criteria discussed in section 2 such as completeness of the model as a description of the physical processes (the primary focus of GCM developments) and accordance with scientific theory and key stylised facts. The reliance on predictive accuracy alone cannot be dominant in the case of climate modelling for the fundamental reason that the application of the GCM models for decadal forecasting is to a domain yet to be observed. But the apparent weaknesses in observed performance on shorter time scales, and more important, the structural weaknesses identified here suggest that a reliance on policy implications from the general circulation models, in particular the primary emphasis on CO_2 is misguided (a conclusion others have reached following a different line of argument, Pielke et al., 2009). Whatever the successes of the decadal forecasting initiative, the resulting forecast uncertainty over policy-relevant time-scales will remain large. The political issue then is to shift the focus from point forecasts to the high levels of uncertainty around them and the need for robust policy responses, a call made by researchers such as Hulme and Dessai (2008) and Pielke (2003); see (Dessai and Hulme, 2004) for an early summary of the debate. The scientific community has surely taken unnecessary risks in raising the stakes so high whilst depending on forecasts and models that have many weaknesses. In particular, they may well fail in forecasting over decades (a period beyond the horizon of most politicians and many voters), despite their underlying explanatory strengths. A more eclectic approach to producing decadal forecasts is surely the way ahead and a research strategy that explicitly recognizes the importance of forecasting and forecast-error analysis.

Acknowledgment: We would like to thank Dr Doug Smith for both supplying us with the data on which this paper is based and helping us translate the language of climatologists into something closer to that of forecasters. Dr Andrew Jarvis has also helped us in the task of interpretation. The remaining infelicities are our own responsibility.

References

Allen, R.J., & Sherwood, S.C. 2008. Warming maximum in the tropical upper troposphere deduced from thermal winds. *Nature Geoscience*, 1: 399 - 403.

Armstrong, J.S. 1985. *Long Range Forecasting: From Crystal Ball to Computer*. New York: Wiley.

Armstrong, J.S. (Ed). 2001. *Principles of Forecasting*. Norwell, Ma: Kluwer.

Armstrong, J.S., & Fildes, R. 2006. Making progress in forecasting. *International Journal of Forecasting*, 22: 433-441.

Beven, K. 2002. Towards a coherent philosophy for modelling the environment. *Proceedings of the Royal Society a-Mathematical Physical and Engineering Sciences*, 458: 2465-2484.

Chatfield, C. 2001. Prediction intervals for time-series forecasting. In: Armstrong, J.S. (Ed). *Principles of Forecasting*. Norwell, Ma: Kluwer.

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M.F., Weber, S., Alcamo, J., Alexeev, V., Berger, A., Calov, R., Ganopolski, A., Goosse, H., Lohmann, G., Lunkeit, F., Mokhov, I., Petoukhov, V., Stone, P., & Wang, Z. 2002. Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Climate Dynamics*, 18: 579-586.

Clements, M.P., & Hendry, D.F. 1995. On the selection of error measures for comparisons among forecasting methods - Reply. *Journal of Forecasting*, 14: 73-75.

Dessai, S., & Hulme, M. 2004. Does climate adaptation policy need probabilities? *Climate Policy*, 4: 107-128.

Dessai, S., & Hulme, M. 2008. How do UK climate scenarios compare with recent observations? *Atmospheric Science Letters*, 9: 189-195.

Douglass, D.H., Christy, J.R., Pearson, B.D., & Singer, S.F. 2007. A comparison of tropical temperature trends with model predictions. *International Journal of Climatology*: n/a.

Fang, Y. 2003. Forecasting combination and encompassing tests. *International Journal of Forecasting*, 19: 87-94.

Fildes, R. 1992. The Evaluation of Extrapolative Forecasting Methods. *International Journal of Forecasting*, 8: 81-98.

Fildes, R., & Ord, J.K. 2002. Forecasting competitions: their role in improving forecasting practice and research. In: Clements, M.P., & Hendry, D.F. (Eds). *A Companion to Economic Forecasting*. Oxford: Blackwell.

Friedman, M. 1953. The methodology of positive economics. In: Friedman, M. (Ed). *Essays in Positive Economics*. Chicago: University of Chicago Press.

Gardner, J.E.S. 2006. Exponential smoothing: The state of the art--Part II. *International Journal of Forecasting*, 22: 637-666.

Granger, C.W.J., & Jeon, Y. 2003. Interactions between large macro models and time series analysis. *International Journal of Finance and Economics*, 8: 1-10.
Granger, C.W.J., & Ramanathan, R. 1984. Improved Methods of Combining Forecasts. *Journal of Forecasting*, 3: 197-204.
Green, K.C., & Armstrong, J.S. 2007. Global warming: Forecasts by scientists versus scientific forecasts. *Energy & Environment*, 18: 997-1021.
Green, K.C., Armstrong, J.S., & Soon, W. 2009. Validity of climate change forecasting for public policy decision making. *International Journal of Forecasting*, 25: 826-832.
Hagedorn, R., Doblas- Reyes, F.J., & Palmer, T.N. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus Series a-Dynamic Meteorology and Oceanography*, 57: 219-233.
Haines, K., Hermanson, L., Liu, C.L., Putt, D., Sutton, R., Iwi, A., & Smith, D. 2009. Decadal climate prediction (project GCEP). *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, 367: 925-937.
Henderson-Sellers, A., & McGuffie, K. 1999. Concepts of good science in climate change modelling. *Climatic Change*, 42: 597-610.
Hulme, M., & Dessai, S. 2008. Negotiating future climates for public policy: a critical assessment of the development of climate scenarios for the UK. *Environmental Science & Policy*, 11: 54-70.
Jose, V.R.R., & Winkler, R.L. 2008. Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24: 163-169.
Kennedy, M.C., & O'Hagan, A. 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 63: 425-450.
Kleindorfer, G.B., O'Neill, L., & Ganeshan, R. 1998. Validation in simulation: Various positions in the philosophy of science. *Management Science*, 44: 1087-1099.
Knutti, R. 2008. Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, 366: 4647-4664.
Kourentzes, N., & Crone, S.F. 2010. Input variable selection for forecasting with neural networks. *Lancaster University Management School*. Lancaster, UK: Lancaster University.
Koutsoyiannis, D., Efstratiadis, A., Mamassis, N., & Christofides, A. 2008. On the credibility of climate predictions. *Hydrological Sciences*, 53: 671-684.
Lakatos, I.i. 1970. Falsification and the methodology of scientific research programmes In: Lakatos, I., & Musgrave, A. (Eds). *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.
Le Treut, H., Somerville, R., Cubasch, U., Ding, Y., Mauritzen, C., Mokssit, A., Peterson, T., & Prather, M. 2007. Historical Overview of Climate Change. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., K.B. Averyt, K.B., Tignor, M., & Miller, H.L. (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 and New York, NY: Cambridge University Press.
Lindzen, R.S. 2009. Global Warming – Sensibilities and Science. *Third International Conference on Climate Change*.
Little, J.D.C. 1970. Models and managers - Concept of a decision calculus
Management Science Series B-Application, 16: B466-B485.
Lopez, A., Tebaldi, C., New, M., Stainforth, D., Allen, M., & Kettleborough, J. 2006. Two approaches to quantifying uncertainty in global temperature changes. *Journal of Climate*, 19: 4785-4796.
Makridakis, S., & Hibon, M. 2000. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16: 451-476.

Meadows, D.H., Meadows, D.L., Randers, J., & Behrens III, W.W. 1972. *The Limits to Growth*. New York: Universe Books (in association with Potomac Associates)
Meehl, G.A., & al., e. 2009. Decadal prediction: can it be skillful? *Bulletin of the American Meteorological Society*.
Meehl, G.A., Stocker, T.F., Collins, W.D., Friedlingstein, P., Gaye, A.T., Gregory, J.M., Kitoh, A., Knutti, R., Murphy, J.M., Noda, A., Raper, S.C.B., Watterson, I.G., Weaver, A.J., & Zhao, Z.-C. 2007. Global climate projections. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., K.B. Averyt, K.B., Tignor, M., & Miller, H.L. (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 and New York, NY: Cambridge U.P.
Oreskes, N. 2003. The role of quantitative models in science. In: Canham, C.D., Cole, J.J., & Lauenroth, W.K. (Eds). *Models in ecosystem science*. Princeton, NJ: Princeton U.P., pp. 13-31.
Oreskes, N., Shraderfrechette, K., & Belitz, K. 1994. Verification, validation, and confirmation of numerical-models in the Earth-sciences *Science*, 263: 641-646.
Parker, W.S. 2006. Understanding pluralism in climate modeling. *Foundations of Science*, 11: 349-368.
Petit, J.R., Jouzel, J., Raynaud, D., Barkov, N.I., Barnola, J.M., Basile, I., Bender, M., Chappellaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V.M., Legrand, M., Lipenkov, V.Y., Lorius, C., Pepin, L., Ritz, C., Saltzman, E., & Stievenard, M. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, 399: 429-436.
Phillips, T.J., Achutarao, K., Bader, D., Covey, C., Doutriaux, C.M., Fiorino, M., Gleckler, P.J., Sperber, K.R., & Taylor, K.E. 2006. Coupled Climate Model Appraisal: A Benchmark for Future Studies. *Eos*, 87: 185,191-192.
Pidd, M. 2003. *Tools for Thinking*. Chichester, UK: Wiley.
Pielke, R.A. 2008. Climate predictions and observations. *Nature Geoscience*, 1: 206-206.
Pielke, R.A., Sr., Davey, C.A., Niyogi, D., Fall, S., Steinweg-Woods, J., Hubbard, K., Lin, X., Cai, M., Lim, Y.-K., Li, H., Nielsen-Gammon, J., Gallo, K., Hale, R., Mahmood, R., Foster, S., McNider, R.T., & Blanken, P. 2007. Unresolved issues with the assessment of multidecadal global land surface temperature trends. *J. Geophys. Res.*, 112.
Pielke, R.A.J. 2003. The role of models in prediction for decisions. In: Canham, C.D., Cole, J.J., & Lauenroth, W.K. (Eds). *Models in ecosystem science*. Princeton, NJ: Princeton U.P., pp. 113-137.
Pielke Sr., R., Beven, K., Brasseur, G., Calvert, J., & M. Chahine, R.D., D. Entekhabi, E. Foufoula-Georgiou, H. Gupta, V. Gupta, W. Krajewski, E. Philip Krider, W. K.M. Lau, J. McDonnell, W. Rossow, J. Schaake, J. Smith, S. Sorooshian, and E. Wood, 2009: Climate change: T, Vol. 90, No. 45, 10 November 2009, 413. 2009. The need to consider human forcings besides greenhouse gases. . *Eos Transactions AGU*, 45: 413.
Randall, D.A., Wood, R.A., Bony, S., Colman, R., T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J., Shukla, J., J. Srinivasan, R.J. Stouffer, Sumi, A., & Taylor, K.E. 2007. Cilmate Models and Their Evaluation. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., K.B. Averyt, K.B., Tignor, M., & Miller, H.L. (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 and NewYork, NY: Cambridge University Press, .
Reichler, T., & Kim, J. 2008. How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, 89: 303-+.
Royer, D.L., Berner, R.A., & Park, J. 2007. Climate sensitivity constrained by CO2 concentrations over the past 420 million years. *Nature*, 446: 530-532.

Shackley, S., Young, P., & Parkinson, S. 1999. Concepts of good science in climate change modelling - Response to A. Henderson-Sellers and K. McGuffie. *Climatic Change*, 42: 611-617.
Shackley, S., Young, P., Parkinson, S., & Wynne, B. 1998. Uncertainty, complexity and concepts of good science in climate change modelling: Are GCMs the best tools? *Climatic Change*, 38: 159-205.
Singer, S.F., & Idso, C. 2009. *Climate Change Reconsidered: The Report of the Nongovernmental International Panel on Climate Change (NIPCC)*. Chicago, IL.: The Heartland Institute.
Smith, D.M., Cusack, S., Colman, A.W., Folland, C.K., Harris, G.R., & Murphy, J.M. 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, 317: 796-799.
Stainforth, D.A., Allen, M.R., Tredger, E.R., & Smith, L.A. 2007. Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, 365: 2145-2161.
Stern, N. 2007. *The Economics of Climate Change: the Stern Review*. Cambridge: Cambridge U.P.
Tebaldi, C., Smith, R.L., Nychka, D., & Mearns, L.O. 2005. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18: 1524-1540.
Trenberth, K. 2007.
Vogl, T.P., Mangis, J.K., Rigler, A.K., Zink, W.T., & Alkon, D.L. 1988. Accelerating the convergence of the backpropagation method. *Biological Cybernetics*, 59: 257-263.
Young, P.C., & Jarvis, A. 2002. Data-Based Mechanistic Modelling, the Global Carbon Cycle and Global Warming. Lancaster University.
Zhang, G.Q., Patuwo, B.E., & Hu, M.Y. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14: 35-62.