Available online at www.sciencedirect.com

SciVerse ScienceDirect

International Journal of Forecasting 27 (2011) 968–995

www.elsevier.com/locate/ijforecast

Validation and forecasting accuracy in models of climate change

Robert Fildes*, Nikolaos Kourentzes

Lancaster Centre for Forecasting, Lancaster University, Department of Management Science, United Kingdom

Abstract

Forecasting researchers, with few exceptions, have ignored the current major forecasting controversy: global warming and the role of climate modelling in resolving this challenging topic. In this paper, we take a forecaster's perspective in reviewing established principles for validating the atmospheric-ocean general circulation models (AOGCMs) used in most climate forecasting, and in particular by the Intergovernmental Panel on Climate Change (IPCC). Such models should reproduce the behaviours characterising key model outputs, such as global and regional temperature changes. We develop various time series models and compare them with forecasts based on one well-established AOGCM from the UK Hadley Centre. Time series models perform strongly, and structural deficiencies in the AOGCM forecasts are identified using encompassing tests. Regional forecasts from various GCMs had even more deficiencies. We conclude that combining standard time series methods with the structure of AOGCMs may result in a higher forecasting accuracy. The methodology described here has implications for improving AOGCMs and for the effectiveness of environmental control policies which are focussed on carbon dioxide emissions alone. Critically, the forecast accuracy in decadal prediction has important consequences for environmental planning, so its improvement through this multiple modelling approach should be a priority.

Keywords: Validation; Long range forecasting; Simulation models; Global circulation models; Neural networks; Environmental modelling; DePreSys; Encompassing; Decadal prediction

1. Introduction

Of all of the areas of forecasting that have succeeded in gaining public attention, the current forecasts of global warming and the effects of human activity on the climate must surely rank amongst the most important. Even before the Kyoto treaty of 1997 there was an emerging scientific consensus on global warming identified with the Intergovernmental Panel on Climate Change (IPCC). By the time of the Fourth Assessment Report in 2007,[1] few scientists working in the field did not accept two central tenets from the IPCC's work: that the earth was warming and that some part of the warming was due to human activity (see Bray & von Storch, 2008). Nevertheless, there have long been powerful counter-voices, both political

* Corresponding author. Tel.: +44 1524 593879.
E-mail address: R.Fildes@lancaster.ac.uk (R. Fildes).

[1] See http://www.ipcc.ch/publications_and_data/publications_and_data.shtml .

0169-2070/$ - see front matter © 2011 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2011.03.008

and scientific, which either denied the first tenet or accepted it but did not accept that human activity was a major causal force. In the political sphere, for example, both the Australian Prime Minister John Howard, in office from 1996 to 2007, and the USA President George W. Bush, from 2001 to 2008, dismissed the notion of global warming. From a scientific perspective, a disbelief in global warming is found in the work of the Heartland Institute and its publications (Singer & Idso, 2009), and supported by the arguments of a number of eminent scientists, some of whom perform research in the field (see Lindzen, 2009). The continuing controversy (see for example Pearce, 2010) raises questions as to why the 4th Report is viewed by many as not providing adequate evidence of global warming. The aims of this discussion paper are to review the various criteria used to appraise the validity of climate models, and in particular the role of forecasting accuracy comparisons, and to provide a forecasting perspective on this important debate which has thus far been dominated by climate modellers. We focus on decadal forecasts (10–20 years ahead). Such forecasts have many policy-relevant implications for areas from land-use and infrastructure planning to insurance, and climatologists have shown an increasing interest in this "new field of study" (Meehl et al., 2009). Decadal forecasts also provide a sufficient data history for standard forecasting approaches to be used.

In Section 2 of this paper, we first set out various viewpoints 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 we pay particular attention here to the role of forecasting benchmarks and forecast encompassing,[2] an aspect neglected by climate modellers generally, as well as by the IPCC Working Group 1 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, we examine the effect on the forecasting performance of including CO_2 emissions and CO_2 concentrations in a nonlinear multivariate neural network that links emissions as an input with global temperatures as an output.[3] 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 which is not yet included in the HadCM3 forecasts. Section 3 also considers disaggregate forecasts of local temperatures.

While our results add further evidence of global warming from a forecasting perspective, there is only limited evidence of a predictive relationship between annual emissions of CO_2 and the 10- and 20-year-ahead global annual average temperature. However, looking to the conclusions, simple forecasting methods apparently provide forecasts which are at least as accurate as the much more complex GCMs for forecasting the global temperature. The last section reflects on the link between the comparative forecasting accuracy and model validation, and its importance in building climate models. Finally, we offer recommendations to the climate-change scientific community as to 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 the details, are all examples of Coupled Atmospheric-Ocean General Circulation Models (AOGCMs).[4] Müller (2010) provides a recent view of their construction and use in both scientific endeavour and policy which is compatible with our own more extended discussion. A brief summary of their basis is as follows. They are systems of partial

[2] Standard forecasting terms are defined in the 'Forecasting dictionary' available at www.forecastingprinciples.com. 'Forecast benchmarks' are forecasts produced by simple models which are regularly used for comparisons with more complicated models. A forecasting method is said to 'forecasting encompass' another if the second set of forecasts adds nothing to the forecast accuracy of the first method.

[3] We also experimented with multivariate networks that used both CO_2 emissions and atmospheric concentrations as inputs.

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

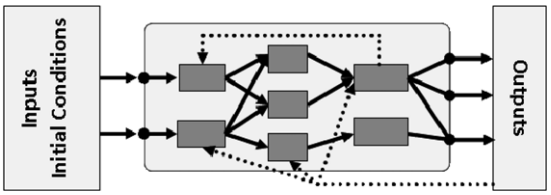

Fig. 1. Stylised representation of Global Circulation Climate Models (GCMs).

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 transfers, radiation, relative humidities, ocean temperatures and flows, and the surface hydrology within each grid cell, then evaluate the interactions with neighboring points. The outputs include 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 the 'average global temperature'. The inputs (termed 'boundary conditions' by climate modelers) include emissions of atmospheric gases (including CO_2) and volcanic eruptions. A crucial intermediate variable is the concentration of CO_2. Fig. 1 shows a stylised representation of such models.

The initial conditions and parameters must be set to solve the partial differential equations at the heart of the model numerically. The initial conditions are fixed, depending on the starting point of the runs, which are often many hundreds of years in the past. At that distance in the past, the observations are limited (from measures such as ice cores), and therefore the starting values are based on plausible assumed pre-industrial states (Meehl et al., 2009). The parameters in the GCM are based on physical (sub)models, which sometimes determine a parameter exactly, while on other occasions the model used is a simplified abstraction. Alternatively, they may be 'tuned' (estimated or calibrated, 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,[5] in climate modeling terms). The aim is to provide a 'best' estimate of the true state of the world's climate system, and corresponding prediction equations both 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 between the observed and simulated outputs. Further tuning is then used to ensure that the model is back on track (e.g., "to deduce the ocean-heat flux convergence field", see Stainforth et al., 2005). In addition, from approximately 1850, observed data on 'forcing', namely exogenous variables (in statistical terminology; known as boundary conditions in climate science), such as CO_2 and volcanic emissions, are included as well. Other potentially relevant variables such as land use changes are usually excluded. Because of the complexity of such models, the computer costs of optimizing these steps are currently prohibitive. Even if it were feasible, given the large number of degrees of freedom and the limited observations, it is necessary to use judgment. Thus, a major part of the model building is judgmental (Stainforth, Allen, Tredger, & Smith, 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 being key to a good model performance. Climate modellers draw a distinction 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" (Haines et al., 2009), and thus, current observations are not usually used to ground (or 'assimilate') the model in the data (although research is currently being conducted in this area). Note that the observed data correspond to 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 rendered compatible with the current observations (through data assimilation; see Mochizuki et al., 2010, for an example). For the IPCC forecasts,[6] this has not been done, since they focus primarily on the

[5] At its simplest, data assimilation combines an estimate of the state of the modelled system with the observed data; the Kalman filter is a simple example. See http://en.wikipedia.org/wiki/Data_assimilation, or, for a more complete explanation of its use in environmental modelling, see Beven (2009).

[6] We use 'IPCC forecasts' as short-hand for the simulated forecasts from AOGCM, conditional on selected scenarios,

longer term. Recently, various modelling exercises have focussed, for reasons which we have already explained, on decadal prediction (Haines et al., 2009; Meehl et al., 2009; Smith et al., 2007). The forecasts from the GCMs use the observations at the forecast origins as their initial values, as we explain in greater detail in Section 3.

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 which are as "realistic as possible" (Beven, 2002). However, various models of sub-systems (e.g. Earth Systems Models of Intermediate Complexity (EMICs)) have been constructed to deliver 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, but with AOGCMs at the extreme.

There is feedback between the outputs and precursor variables, with varying, often long, lags and nonlinearities; for example, Young and Jarvis (2002) show that there is nonlinear temperature-driven feedback operating on the intermediate relationship between CO_2 emissions and atmospheric CO_2. When allied to the nonlinear effects of atmospheric CO_2 on radiative forcing, one would anticipate that the control relationship of interest between CO_2 emissions and temperature, through the intermediate variable, CO_2 concentrations, is likely to be nonlinear (though possibly nearly linear over some input domains). Long 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; see for example Beven (2002, 2009), Kennedy and O'Hagan (2001) 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 to 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 for model parameters to be determined uniquely from theoretical considerations. Instead, they will depend on many factors, including the specific location where they are applied (Beven, 2002, Section 3; see also Beven, 2009). Nor does the problem disappear with increased disaggregation; indeed, Beven argues that increased disaggregation may make matters worse.

  The parameters in a GCM are sometimes 'tuned', but are rarely optimally estimated. When a set of parameters is estimated, they are likely to suffer from the standard problem of multicollinearity, or more generally non-identifiability, due to the models being over-parameterised (unless the physics of the problem can be used to identify the parameters). A key point to note is that possible nonlinear effects (e.g. the CO_2 absorption capacity of a forest at levels of atmospheric CO_2 twice that currently observed) cannot be known or reliably estimated. As Sundberg (2007) points out in an empirical study of climate modellers, there is a considerable degree of argument as to how GCMs should be parameterised.

(iii) Uncertainty arising from model misspecification.
- For example, in the current generation of AOGCMs, certain potentially important processes such as cloud effects and water vapour

produced by various modelling agencies and discussed in the IPCC assessment reports. There is a considerable degree of confusion in regard to terminology within the GCM community, with the term 'projection' being used in an attempt to avoid the issue of accuracy. See for example the discussion by Pielke Sr. (2005).

formation are still poorly understood. A second example is the way in which vegetation is modelled. Aggregation over time and space also leads to misspecification. However, a greater disaggregation does not lead to a better-specified model, as Beven (2009) has explained, since it leads to the inclusion of non-identifiable parameters. A necessary consequence of parameter uncertainty and specification uncertainty is that the limits of acceptability of the set of models (in model space, in the terminology of Beven, 2002) that represent the global climate might need to be greater than observational error would suggest. Therefore, a model should not necessarily be rejected in a "relaxed form of Popperian falsification" when it is incompatible with the observations (Beven, 2002); all models fail in some important attributes. Despite the fact that this is the common view, Knutti (2008) claims that they all offer "credible approximations to the descriptions of the climate system given our limited understanding". In contrast, a survey within the climate science communities showed that there is a diversity of views, only some of which can be described as being supported by a majority of scientists (see Bray & von Storch, 2008). Thus, model misspecification remains a serious issue (as we will show).

(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. Critically, however, the observed world is stochastic, not least because of the actions of actors in the system (see Koutsoyiannis, 2010, for an exploration of this issue).

(v) Uncertainty in the data.
- There remains considerable degree of controversy as to the choice of measure for the key variable, temperature, whether at an aggregate level or 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 Sr. et al., 2007).

and

(vi) Numerical and coding errors.
- In the solution to the system equations, both unavoidable numerical errors and coding errors ('bugs') may occur.

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.

Various approaches for mitigating these uncertainties have been proposed. Ensemble methods provide a combined set of predictions (Hagedorn, Doblas-Reyes, & Palmer, 2005), which may be based on runs from different initial conditions. In addition, some aspects of the specification uncertainty are alleviated through multi-model averaging. The results from comparing the benefits of the two approaches to alleviating uncertainty for within-year seasonal forecasting show 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 previously been noted in the discussions on climate.

There is currently debate as to appropriate methods of model averaging (Lopez et al., 2006). A Bayesian approach (Tebaldi, Smith, Nychka, & Mearns, 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 super population of AOGCMs). This leads to either uni- or multi-modal probability density functions, where the latter are the result of the models disagreeing. Substantially different results arise from these different methods. As yet there is no reason to believe that the conclusion of this debate will depart from that in the forecasting literature, namely recommending a simple or trimmed average for the most accurate point forecast (Jose & Winkler, 2008). The range of forecasts from a selected group of GCMs or the estimated probability density function of the ensemble offers 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 a common modelling paradigm and methods, data and the limitations imposed by current computer hardware. Counterintuitive forecasts that do not fit with the consensus are either given a low weight (as in the Bayesian combination) or omitted (for example, if a new ice age is foreseen; see Beven, 2002).

The effects of uncertainty in the forcing variables are dealt with primarily through the use of policy scenarios that aim to encompass the range of outcomes so as to guide policy and decision making (Dessai & Hulme, 2008). When 'hindcasting', the term used by climate modellers to describe conditional forecasting, this approach may leave out known events such as volcanic eruptions (e.g. the Mt. Pinatubo eruption in 1991) from the simulated future path. Alternatively, including such stochastic interventions in the simulation can give an estimated distribution of future outcomes, conditional on the particular emissions scenario.

The uncertainty in a forecast is usually measured through a predictive probability density function. In the forecasting literature, the various model-based methods for estimating the future error distribution (see Chatfield, 2001) are all (often necessary) substitutes for observing the error distribution directly through an out-of-sample evaluation or 'hindcasting'. In general, it is likely that none of the model-based estimates of the predictive density function (and prediction intervals) will be any better calibrated in climate forecasting than in other applications (Stainforth et al., 2007). The importance of examining the empirical error distribution has been recognized in principle by the IPCC, although, as Pielke Jr. (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, partly because of the computational complexity of GCMs.

For long horizons (100+ years), climate modellers have tended to dismiss the prospect of estimating the conditional forecast error distribution, arguing that models of the effects of slower physical processes such as the carbon cycle rely on proxy data (e.g. ice records) which have been used in the model construction. This effectively renders the comparison between the model forecasts and the observations an 'in-sample' test, in that the models have been refined to match the historical record. Such a comparison can be no more than weakly confirmatory (Stainforth et al., 2007).

In summary, while all of the authors we have referred to recognize the match between model predictions and their associated prediction intervals as a key criterion for appraising the different GCMs, few, if any, studies have made a formal examination of their comparative forecasting accuracy records, 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? An early attempt to clarify the difference was given by Armstrong (1985), who points out the difficulty of a clear definition, but suggests that what distinguishes long term forecasting is the prospect of large environmental change. Curiously, the book *Principles of Forecasting* (Armstrong, 2001), which aims to cover all aspects of forecasting, pays no particular attention to the topic, apart from a similar definition, regarding the forecasting approaches covered within as applicable. In climate modelling and forecasting,[7] we have already seen a dramatic change in the forcing variable of CO_2 emissions over the past 150 years, leading to concentration levels which have not been seen for thousands of years, with scenarios predicting a doubling over the next 50 years,[8] leading to a further 2.0–5.4 °C increase in this century in the high-emissions IPCC scenario (A2). Thus, the condition of dramatic exogenous environmental change is expected.

We suggest that the main reason why this is important for validation when large changes are expected is that any forecasting model designed to link CO_2 emissions (or any other induced forcings such as changed land use) with temperature changes must aim to establish a robust relationship between the two in the future, not yet observed, world and not just in the past. Thus, the standard approaches to validation which are adopted in the forecasting literature (Armstrong, 2001) are not sufficient in themselves.

7 http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-spm.pdf.

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

Oreskes, Shraderfrechette, and Belitz (1994), marshalling the logic of the philosophy of science, have argued that open system models such as AOGCMs cannot be verified; only certain elements of a model, such as the numerical accuracy of its forecasts, can be. Nor can they be validated in the strongest sense of the word, implying the veracity of the model under review. While some climate modellers with a forecasting orientation[9] have perhaps taken the view that a valid model should realistically represent the 'real' system in depth and detail, forecasting researchers, in contrast, have taken a more comparative view of validity. From a forecasting perspective, GCMs can be used to produce out-of-sample ex post forecasts ('hindcasts'), conditional on a particular set of forcing variables (such as emissions) or an intermediate variable (such as atmospheric gas concentrations). The ex post forecasts also depend on data which would have been available to the modeller at the forecast origin. (Of course, the model should not be modified in the light of the out-of-sample data in order to produce better 'forecasts'; however, this seems unlikely to be a problem with GCMs because of their complexity.) To forecasting researchers, the validation of a model using ex post errors 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 under review produces more accurate forecasts than alternative forecasting models. The match between the known physical characteristics of the system and the model is seen as less important. Forecasting models (like all simulation models) are seen as being valid only temporarily, designed for particular uses and users, and subject to repeated confrontations with the accumulating data (Kleindorfer, O'Neill, & Ganeshan, 1998).

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

In summary, the claim that as "realistic (a model) as possible" (Beven, 2002) will necessarily produce the most accurate forecasts has long been falsified within forecasting research; for example, Ascher (1981) considered a number of application areas, including energy modelling and macroeconomic forecasting, and criticised such large macro models for their inadequate forecasting accuracy. More recently Granger and Jeon (2003) revisited the argument that small (often simple) models are the most effective. In fact, Young and Parkinson (2002) showed that simple stochastic component models can emulate the outputs of much more complex models by identifying the dominant modes of the more complex model's behaviour. Thus, with the focus being on the forecasting accuracy and its policy implications, the requirement for valid models (and forecasts) requires the construction of an accuracy record, which, in principle, could be done with GCMs.

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

[9] While some climate modellers have been concerned with sub-system interactions and necessarily adopt a heavily disaggregated modelling approach, the GCMs more often have a major forecasting focus.

2.2. Climate forecasting—defining the problem context

"All models are incorrect, but some are useful".[10] Any meaningful evaluation must specify (i) the key variables(s) of interest, such as the 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.

With regard to specifying the variable(s) of interest and the forecast horizon, while a substantial degree of attention has been paid to the aggregate forecasts, particularly those of temperature, the AOGCM forecasts are highly disaggregate and use increasingly small spatial grids. Their corresponding localised forecasts of temperature, precipitation and extreme events have been publicized extensively 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 one decade to centuries ahead. In particular, they are not intended as short-term forecasting tools, although Randall et al. (2007), in the IPCC report, 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 preceding paragraphs, models which are 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; here, we have chosen a 10–20 year horizon, which is short from a climate modelling perspective. It is, however, relevant to infrastructure upgrades, energy policy, insurance, etc., and, as noted, has increasingly become the focus of at least 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 in the past 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 on the validity of the models for decadal forecasting, CO_2 emissions can be regarded as known, at least for 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 the 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 for 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 for models to be valuable to their users, they 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 [feasible] outputs". Various modellers concerned with environmental policy have also examined the role of models. For example, Pielke Jr. (2003) proposes guidelines that support and extend the work of Little, with particular emphasis on 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). However, there has been less attention paid to the remaining criteria. With such a wide range of stakeholders, the IPCC have chosen to present their models to 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) being kept at a distance. Although, in principle, the models are comprehensible, their completeness (and complexity) means that there has been relatively little experimentation aimed at testing 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 "exploring predictability" and get closer to Little's requirements.

[10] Usually attributed to George Box.

2.3. 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) using 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, Meadows, Randers, & Behrens, 1972), which, whilst much more aggregated than the current generation of AOGCMs, included additional variables measuring population, technology and the economy, as well as environmental variables. Though it was intended primarily as a policy tool, the 'Limits' authors inevitably slipped back into forecasts (conditional on various policies). In this early world modelling exercise, no attempt was made to demonstrate that the model had any forecasting abilities when compared to alternative methods.

As part of the early debate on economic model building, Friedman (1953) placed predictive ability at the head of his list of requirements for 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 should therefore be evaluated by 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 as either known or unknown, depending on the purpose of the model evaluation.) If one model is to be preferred over another (based on this criterion), then the observed errors on past data should be smaller (for the relevant measures, e.g. Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), or turning point predictions). One fundamental contribution of forecasting research is its emphasis on the requirement that a method (or forecasting process) demonstrates its superiority by beating some plausible competing benchmark. In so far as researchers know how to select a good forecasting method ex ante, perhaps the primary requirement is that it must have been shown to work previously in circumstances similar to those which are expected to apply in the future, outperforming the alternatives, and in particular a benchmark (Armstrong & Fildes, 2006). Of course, it is expected that in small samples, the noise may well overwhelm the signal (in the GCMs derived from increasing CO_2 emissions and concentration levels), and therefore a large sample of forecasts may need to be considered.

A number of researchers have criticised the IPCC models and forecasts for their failure to provide any evidence of their predictive accuracy, despite the IPCC's strong claims (Green & Armstrong, 2007; Pielke Sr., 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 conditional forecasts from these climate models, based on various anthropogenic scenarios, aim to induce novel (and potentially expensive, see for example Stern, 2007) policies, the importance of the IPCC models delivering ex post forecasts which are more accurate than the competing alternatives cannot be overestimated. Reliable prediction intervals are also needed. In addition, localised forecasts derived from the AOGCMs need to be subjected to the same tests, since policies will typically be implemented locally (see for example Anagnostopoulos, Koutsoyiannis, Christofides, Efstratiadis, & Mamassis, 2010, and Koutsoyiannis, Efstratiadis, Mamassis, & Christofides, 2008; 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 which take dependencies into account need to be constructed (see Reichler & Kim, 2008, or, within the forecasting literature, Clements & Hendry, 1995).

The forcing (exogenous) variables are measured with error, and features such as major volcanic eruptions may produce large errors in some models (perhaps because of dynamic effects) 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 forecasts do not pose any fundamental issues that earlier discussions of forecast evaluation have not considered. However, the size of these models apparently discourages the obvious resolution of this problem:

fix a starting date where the exogenous variables are regarded as being measured reliably (within some range), 'tune' the model to match the in-sample data, and calculate the out-of-sample rolling origin forecast errors.[11] 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 the model stability, the variability of the model output compared with the observed behaviour, and the consistency with observations, which are 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 practical in principle.

In summary, there is an increased recognition within the climate modelling community of the importance of forecasting accuracy, with a focus on decadal prediction. This is leading to a greater emphasis on data assimilation methods for initialising the forecasts if effective forecasts are to be produced (Mochizuki et al., 2010; see also http://www.clivar.org/organization/decadal/decadal.php).

2.4. Stylised facts

A second aspect of validating a forecasting model is the need for models which are capable of capturing the stylised facts of climate fluctuations. The term 'stylised fact' here is used conventionally[12] to mean a simplified characterisation of an empirical finding. Here, the GCMs aim to simulate various stylised facts in the current climate record, and potentially the more distant past as well. Such stylised facts include the changing temperature trend over the last century, the effects of major volcanic eruptions and the cyclical effects of the El Niño-Southern Oscillation phenomenon, for example. This criterion applies with additional force when either there is no 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 the model outputs, which inevitably reveal some (possibly temporary) discrepancies between the model outputs and the observed behaviour.

2.5. Black-box and white-box validation

Because the GCMs are intended for use beyond the range of some of their input variables (most critically, emissions) and expected outputs (e.g. temperature), other validation criteria beyond comparative forecast accuracy come into play. These are needed to enable us to understand and model the input-output relationships between the variables which are seen as primary causal inputs (and in particular emissions, as they affect system outputs such as temperature and precipitation). Pidd (2003) remarks that "(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, which is theoretically justified by science-based flow models, as shown in the system in Fig. 1 (e.g., emissions and atmospheric CO_2).

The GCMs 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 the corresponding temperature observations cannot include emissions at double the current level). Thus, it is important for the models to demonstrate robust and plausible dynamic responses to inputs outside the observed range. The 'Climate prediction.net' experiment has been used to deliver some evidence on the both model and initial condition sensitivity to a doubling of CO_2 (Stainforth et al., 2005), with the results showing extremes of response (even including cooling). The experiment has also been used to examine the joint parameter sensitivity, compared to the effects of single parameter tests. The former are needed, as here, because the overall effects may be more than the sum of the individual sensitivities.

Intensive research in analysing sub-systems of the GCMs continues to be carried out at both the local and

[11] The deterministic nature of the models makes the rolling origin requirement more relevant because of the effects of the initial conditions at the forecast origin.

[12] See http://en.wikipedia.org/wiki/Stylized_fact.

regional levels, but also including, for example, the flow relationships between the land, atmosphere and ocean. The logical next step is to add open box support to the global models.

2.6. Process validation

The scientific community has developed its own procedures for assessing the validity of the models it develops. They depend primarily on peer review and replicability through open access to the proposed models and computer code, the data on which they are based and the models' outputs. At the heart of the processes is the concept of falsifiability (Popper, 2002; but see Kleindorfer et al., 1998 and Oreskes et al., 1994, for a more focussed discussion in relation to GCMs) through critical predictive tests and replicability. Openness in making both the data and models available is at the heart of both criteria. However, the peer review process acts as a limiting gateway to researchers from outside the mainstream climate community wishing to gain 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 which are regarded as worthy of development (Shackley, Young, Parkinson, & Wynne, 1998). Unfortunately, the existence of a scientific consensus is no guarantee of validity in itself (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 was addressed in an exchange between Henderson-Sellers and McGuffie (1999) and Shackley, Young, and Parkinson (1999), with the latter arguing that, despite different styles of modelling, the predominant approach is 'deterministic reductionist'; that is to say, the GCMs as described here (rather than, for example, aggregate statistical). More recently, Koutsoyiannis (2010) has argued for a stochastic approach to complement the deterministic reductionist GCM approach. Pearce (2010) also gives some insights into the tensions within the community of climate scientists that may have led to hostility to critics outside the dominant GCM community. However, no critique of the GCM approach has yet become established, either inside or outside the global climate-modelling community.

2.7. Climate scientists' viewpoints on model validation

The IPCC Report contains the most authoritative views of climate scientists on model validation, often with a detailed discussion of the issues raised above (Le Treut et al., 2007). The IPCC authors recognize all of these elements of model validation, and summarise 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 that "less credence should be given to the assertion until it is tested and independently verified". The perspective which the authors adopt is one where cumulative evidence of all of the types discussed above is collected in order to discriminate between one model (or explanation) and another, whilst accepting a pluralistic (multi-model) perspective as reasonable practice (Parker, 2006). This is wholly compatible with the long-established but unacknowledged literature on the implications of the philosophical foundations of simulation 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, Section 8.1.2.3), has not taken 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". In fact, the report backs away from model comparison and criticism, arguing that 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. However, the precise

meaning of this and many similar statements is far from clear. The models themselves differ quite substantially on such key parameters as climate sensitivity (Kiehl, 2007; Parker, 2006) and the incorporation of aerosol emissions.

Chapter 8 of the report 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. Moreover, relationships that embrace a broader set of possible anthropogenic forcing variables are not represented by the models included in the report (Pielke Sr., 2008). A related issue, although one which does not itself deliver 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 eight EMIC models listed in Chapter 10, *Global climate projections* (Meehl et al., 2007),[13] the models have apparently not been used for forecast comparisons.

The discussion on model validation in the climate modeling community has moved on somewhat since the IPCC report of 2007, with a greater emphasis on the conformity of models with observations. Quite recently, research programs have been developed by climate modelers for comparing models (e.g., the Program for Climate Model Diagnosis and Intercomparison, Phillips et al., 2006) and examining forecasting accuracies (Haines et al., 2009; Keenlyside, Latif, Jungclaus, Kornblueh, & Roeckner, 2008; Smith et al., 2007). The results from comparing models have shown that a combination of forecasts from different models is more effective than a single model (see for example Hagedorn et al., 2005), and that the improvement as a result of 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 could potentially inform us as to where improvements might be possible, although such an appraisal has not yet been done (to the best of the authors' knowledge).

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 the GCM community, rests primarily on the physical science of the sub-models, rather than on their predictive abilities. The models also capture the stylised facts of climate such as the El Niño and the Southern Oscillation. While the IPCC authors note that there is a considerable degree of agreement between the outputs of the various models, the forecasts do differ quite substantially, and the combined model forecasts apparently conform to recent data better than any single model. The omissions in Chapter 10 of the IPCC report and most of the subsequent research lie in the lack of evidence that the models actually 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 tempered somewhat by the recognition that the physical sub-models are supposedly robust over the increasing CO_2 emissions input, and key experimental parameters in the physical laws embedded in the models should remain constant. In fact, climate modellers have raised 'completeness' in model building above all other criteria when evaluating the model validity. It is not a criterion that earlier simulation modellers have ever regarded as dominant (Kleindorfer et al., 1998); rather, it has often been regarded as a diversion that detracts from both understanding and forecast accuracy.

2.8. Outstanding model validation issues

Despite the siren voices that urge us to reject the proposition that models can be useful 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 which is valuable for policy and decision making.

As forecasters examining the evidence, we have been struck by the vigour with which various stylized facts and the 'white-box' analysis of sub-models are debated. An interesting example is that of tropospheric temperatures: Douglass, Christy, Pearson, and Singer (2007) highlighted a major discrepancy with model predictions, following which Allen and Sherwood (2008) critiqued their conclusions via a web discussion contesting the proposed resolution (see also Pearce, 2010, Chapter 10). Where the debate has been most lacking is in

[13] The keyword search used was 'model name + forecast* + valid*' in Google Scholar.

the emphasis and evidence on the 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 forecasts, 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 (relative to other model-based forecasts), it is surely hard to argue that policy recommendations from such models should be acted upon. The study of the 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 so many possibilities of confusion, why, we might wonder, has the climate change movement gained so much ground, despite entrenched and powerful opposition? From a long-term perspective, there has been a considerable degree of variability in the Earth's climate, both locally and globally. An examination of the ice-core record of Antarctic temperatures suggests a range of 10 °C over the past 400,000 years, as can be seen in Fig. 2. However, changes of more than 2 °C in a century have only been observed once, five centuries ago in what is admittedly local data. Is the observed (but recent) upward trend shown in Fig. 3 nothing more than an example of the natural variability long observed, as argued by Green, Armstrong, and Soon (2009), or is the projected temperature change in the IPCC report exceptional?

For the annual data needed for decadal modelling, there are many time series data sets of aggregate world temperatures, but it is only since 1850 that broadly reliable data have been being collected regularly; the Hadley Centre data series HadCRUT3v is the latest version of a well-established and analysed series used to appraise AOGCMs. More recently, NASA[14] produced alternative estimates which have a correlation of 0.984 with the HadCRUT3v annual dataset (data: 1880–2007). 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 HadCRUT3v data for model building. In making this choice, we pass over the question of whether this series offers an accurate and unbiased estimate of global temperatures. While the resolution of this uncertainty is of primary importance in establishing the magnitude and direction of temperature change, it has no direct effect on our methodological arguments. Fig. 3 shows a graph of the HadCRUT3v data, together with a 10-year centred moving average.

The features of the global temperature time series (the stylised facts) are relatively stable between 1850 and 1920; there is then a rapid increase until 1940, followed by a period of stability until 1970, since which time 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 Fig. 3, ±0.6 °C) have often been broken, but the evidence which we invoke here is local rather than global. We can conclude, however, 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 2.4 in relation to 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 error measures).[15] Therefore, over the time scales which we are concerned with, a forecasting model should allow the possibility of a local trend if it is to capture this particular feature of the data. Of course, if no trend

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

[15] Perhaps some of the scepticism as to global warming is due to the failure of the IPCC to clearly demonstrate such success. Of course, there are 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'.

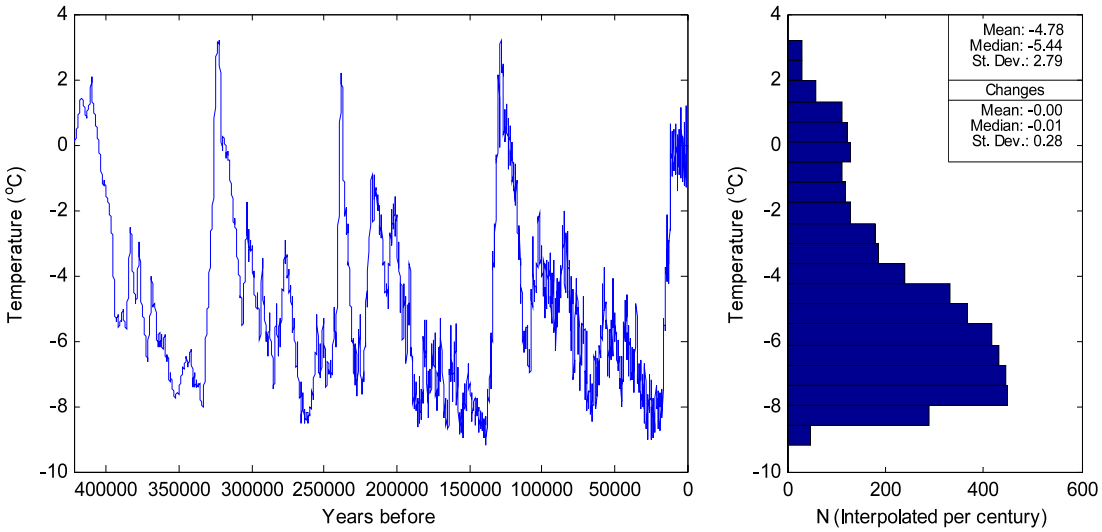

Fig. 2. Vostok ice core temperature estimate plot and histogram. Missing histogram values are interpolated at century intervals in order to provide time equidistant estimations.
Source: Data taken from Petit et al. (1999).

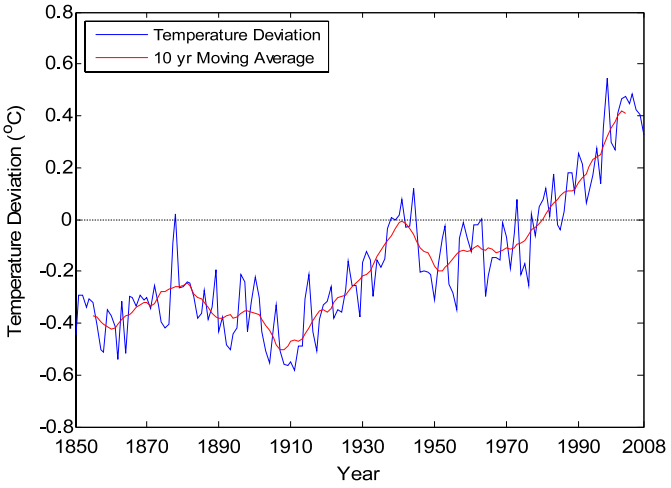

Fig. 3. Temperature anomaly in °C (deviations from the 30 year average temperature, 1961–1990) and a ten year moving average.
Source: Data taken from http://www.cru.uea.ac.uk/cru/info/warming/gtc2008.csv.

is found on the time scale under consideration, this should also emerge from the modelling.

The evaluation of the forecasts produced by GCMs requires a time series history, but this is not straightforward, since there is no established, definitive, long historical record of forecasts. However, we are able to benefit from Smith et al.'s (2007) work, which provides us with a 25-year history of out-of-sample forecasts. While this is only one particular example of a GCM being used in forecasting, it has the (to our knowledge unique) advantage of generating a set of forecasts in much the same way as a forecaster would. Smith et al. used a "newly developed Decadal Climate Prediction System (DePreSys), based on the Hadley Centre Coupled Model, version 3e (HadCM3)", which was 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 using pre-industrial levels of greenhouse gases as inputs until it reaches a 'steady climatological state'—the control run. Most parameters (including constants) are fixed, either theoretically or experimentally. A number of parameters describe processes which 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, though they have long memory.
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 the natural climate variability).
3. The model is then 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 in order to produce quarterly forecasts up to 10 years ahead, with forecasts based on observed forcings (with volcanic forcings only being included once they have occurred).
5. Smith et al.'s final annual forecasts are calculated by averaging across the quarterly forecasts. For one-step-ahead annual predictions, quarterly forecasts from the two preceding years are used, giving an ensemble size of eight members: two quarterly forecasts for each quarter of the year in question. For longer lead times, this is extended further to include the four preceding years, increasing the number of ensemble members to 16. In practice, each annual forecast is the result of a moving average of several years. This only permits the calculation of forecasts up to 9 years ahead.

A partial technical description is given in the online supporting material (see Smith et al., 2007). In the calculations we report below, we use the more straightforward calculation of averaging the four quarterly forecasts, omitting step 5. This allows us a full 10-year-ahead sample. We note that implementing step 5 leads to an improvement in Smith et al.'s forecast errors, particularly for short horizons. Further details are available on the web site for this article at www.forecasters.org/ijf/.

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 10-year-ahead forecasts. Relative to the forecasts produced by HadCM3, which did not take into account the observed state of the atmosphere and ocean, the results (unsurprisingly) were substantially better, as Smith et al. (2007) demonstrate.

The forecasts from the DePreSys model permit a 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) or not. If a trending univariate benchmark is measured to be more accurate ex ante than the naïve no-change benchmark argued for by Green and Armstrong (2007) amongst others, this supports the notion of global warming. (Of course, it tells us nothing about either its causes or possible effective policy responses.)

The DePreSys forecasts are conditional forecasts based on various anthropogenic variables, and CO_2 concentrations in particular. Using annual emissions from 1850 to 2006[16] (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 using the DePreSys forecasts and the univariate benchmarks. This gives us the potential to discriminate between the various effects embodied in the different benchmark models, thus pointing the way to possible improvements in the Hadley GCM model. The various modelling comparisons also give some information on whether CO_2 emissions can be said to Granger-cause global temperatures.

3.1. Evaluating alternative benchmarks

The results of past forecasting competitions provide empirical evidence on the comparative accuracies of various benchmark forecasting methods (Fildes & Ord, 2002; Makridakis & Hibon, 2000), from which

[16] Global fossil fuel CO_2 emissions, total carbon emissions from fossil-fuels (million metric tons of CO_2), http://cdiac.ornl.gov/trends/emis/tre_glob.html.

we will choose some strong performers to consider further here. In addition, we include both a univariate and a multivariate nonlinear neural net. The data used for model building are the annualised HadCrut3v and total carbon emissions from fossil fuels between 1850 and the forecast origin. We consider a number of forecast origins between 1938 and 2006. The estimation sample was extended forward with each new forecast origin and the models were re-estimated. Forecast horizons from 1 to 20 were considered, and were then separated into short- and long-term forecasts.

The random walk (naïve) model 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) and Green et al. (2009) have provided arguments for its use in climate forecasting, although we do not regard as strong over the forecast horizons we are considering here (10–20 years). In addition, we will also try a number of benchmarks which have performed better than the naïve in the various competitions: simple exponential smoothing, Holt's linear trend and the damped trend (Gardner, 2006). The last two incorporate the key stylised fact of a changing local trend. They have been estimated in MatLab® using standard built-in optimisation routines. The 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.[17] These methods are all estimated on the time series of temperature anomaly changes. The multi-step-ahead forecasts are produced iteratively, i.e., the one-step-ahead forecasted value is used as an input in producing the two-step-ahead value, and so on.

In addition, we have also considered both a univariate and a multivariate neural network model (NN). Unlike the other models, these models have the potential to capture nonlinearities in the data, although they are not readily interpretable in terms of the physical processes of the climate system. Furthermore, NNs are flexible models which do not require the explicit modelling of the underlying data structure, a useful characteristic in complicated forecasting tasks such as this one. Nor do they rely on particular data assumptions. The univariate NN is modelled on the differenced data because of non-stationarity, and the inputs are specified using backward dynamic regression,[18] 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. There is no generally accepted methodology for specifying the number of hidden nodes H in the layer (Zhang, Patuwo, & Hu, 1998), and 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 multivariate NNs respectively. Formally, the model is,

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

where $g(x) = \tanh(x) \cong \frac{2}{(1+e^{-2x})-1}$ (Vogl, Mangis, Rigler, Zink, & Alkon, 1988); where $\boldsymbol{X} = [x_1, \ldots, x_I]$ is the vector of I inputs, including lagged observations of the time series and any explanatory variables. The network weights are $\boldsymbol{w} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_H]$ and $\boldsymbol{\gamma} = [\gamma_{11}, \gamma_{12}, \ldots, \gamma_{HI}]$ for the output and the hidden layer respectively. β_0 and γ_{i0} are the biases of each neuron. The hyperbolic tangent activation function $g(\cdot)$ 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 in order to avoid overfitting. 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, in order to mitigate the problems that arise due to the stochastic nature of the NNs' training. The final forecast is calculated as the median output of these 20 different initialisations. The median is used to provide robust forecasts to the different

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

[18] A regression model is fitted and the significant lags are used as inputs to the neural network (Kourentzes & Crone, 2010).

Table 1
Mean and median absolute errors (MAE and MdAE) for forecasting 1–4 years ahead. Average global temperature deviations using alternative univariate and multivariate forecasting methods, compared to Smith et al.'s GCM forecasts from DePreSys. The most accurate method(s) are shown in bold.

MAEs (MdAEs) for forecasting 1–4 years ahead				
	Method	Hold-out sample period		
		1939–2007	1959–2007	1983–2005
Horizon 1–4	Naïve	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 (DePreSys)	–	–	**0.067** (**0.048**)
No. of observations		66	46	20

training initialisations. Finally, the NNs are retrained at each origin. We have used a black-box input-output approach for the multivariate neural nets, using CO_2 annual emissions and lagged values of the temperature anomaly as inputs. Volcanic emissions have been excluded, ensuring that the results are comparable to Smith et al.'s.

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

The primary forecast horizon is the 10- and 20-year-ahead temperature deviation, with the absolute error as the corresponding error measure. However, the compatibility between the shorter-term forecasts (we will consider 1–4 years) and the longer horizon forecasts 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 that use CO_2 emissions, and the combination of forecasts, as well as for different hold-out samples. They are compared to Smith et al.'s forecasts where possible (recall that we have used the raw rather than the moving average forecasts from Smith et al.).

The short-term forecasts show a high variability in the performances of the various extrapolative models. Thus, the combined forecast performs well. The NNs perform well on the longer data set, but the more consistent upward trend over the last 20 years has allowed Holt's local linear trend model to beat them.[19] The forecasts from DePreSys outperformed the statistical models for the shorter hold-out sample period, thus failing to support the view that the GCMs are unable to capture short-term fluctuations. (We note that the moving average process applied by Smith et al. improves the accuracy further.)

3.1.2. Longer-term forecasts

Table 2 shows the results for similar comparisons for the 10- and 20-year-ahead forecasts. Where a comparison with the results of Smith et al. 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 performances. The neural networks and the combined forecasts performed the best overall when evaluated over long hold-out periods. Holt's model outperforms the rest during the period 1983–2005 when there is a significant trend in the data.

While there are no 20-year-ahead forecasts for DePreSys, the multivariate NN that considers CO_2 information consistently performs the best in long

[19] Multivariate NNs that use both CO_2 emissions and atmospheric concentration demonstrate similar performances, with MAEs (MdAEs) of 0.104 (0.088), 0.101 (0.088) and 0.088 (0.70) for the periods 1939–2007, 1959–2007 and 1983–2005, respectively.

Table 2
Mean and median absolute errors (MAEs and MdAEs) for forecasting 10 and 20 years ahead. Average global temperature deviations using alternative univariate and multivariate forecasting methods, compared to Smith et al.'s GCM forecasts from DePreSys.

Method	MAEs (MdAEs) for forecasting 10 and 20 years ahead					
	Hold-out sample period					
	Horizon 10			Horizon 20		
	1948–2007	1968–2007	1992–2007	1958–2007	1978–2007	2002–2007
Naïve	0.152 (0.142)	0.155 (0.142)	0.202 (0.198)	0.202 (0.181)	0.273 (0.276)	0.386 (0.413)
Single ES	0.156 (0.130)	0.168 (0.160)	0.220 (0.242)	0.208 (0.182)	0.290 (0.310)	0.406 (0.404)
Holt ES	0.184 (0.146)	0.136 (0.125)	**0.088** (0.084)	0.355 (0.301)	0.306 (0.284)	0.195 (0.251)
Damped trend ES	0.158 (0.134)	0.161 (0.145)	0.195 (0.189)	0.230 (0.192)	0.287 (0.315)	0.402 (0.406)
AR	0.140 (0.122)	0.131 (0.119)	0.169 (0.156)	**0. 178** (**0. 134**)	0.220 (0.207)	0.312 (0.344)
NN-univariate	0.136 (0.091)	**0.106** (**0.087**)	0.098 (0.079)	0.200 (0.146)	0.175 (0.139)	0.203 (0.210)
NN-multivariate	0.154 (0.136)	0.131 (0.099)	**0.088** (**0. 058**)	0.195 (0.149)	**0.131** (**0. 103**)	**0.125** (**0. 111**)
Combination	**0.133** (**0.113**)	0.118 (0.110)	0.133 (0.131)	0.194 (0.181)	0.212 (0.235)	0.267 (0.273)
Smith (DePreSys)	–	–	0.127 (0.127)	–	–	–
No. of observations	60	40	16	50	30	6

term forecasting over a sample of the last 30 years in the holdout sample. This effect becomes more apparent during the last decade, where the errors of the multivariate NN become substantially lower than those of all of the other models.[20]

Assessing the direction of the errors, all models except for the NNs consistently under-forecast for all periods examined above. On the other hand, Smith et al.'s DePreSys over-forecasts. NNs show the lowest biases, and do not consistently under- or over-forecast.

The unconditional forecasts for the 10- and 20-year-ahead world annual temperature deviations are 0.1 °C–0.2 °C per decade for the methods which are able to capture trends, compared with the best estimate from the various global climate models of 0.2 °C (approximately) for the A2 emissions scenario. The forecasts for all models are provided in Fig. 4, and a summary is given in Table 3. Note that the models which have proved accurate at predicting global temperatures in our comparisons in Table 2, forecast temperature increases for the next two decades (details are given in the paper's supplementary material). The NN-multivariate model provides the same per year temperature increase forecast as the A2 scenario[21] from the IPCC AR4 report.

However, the above analysis does not say anything about the causes of the trend (or even anything much about global warming). Nevertheless, it does show the trend continuing over the next ten or twenty years. It is also quite persistent, in that the full data history shows that there are relatively few rapid reversals of trend. By plotting the changes in the trend component of the 10-year-ahead Holt's forecasts, in Fig. 5, we can observe that the trend estimate remains relatively low and there are very few years with negative trends.

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:

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

$$Temp_t = \alpha_0 + \alpha_1 ForMeth1_{t-h}(h) + \alpha_2 ForMeth2_{t-h}(h) + e_t \quad (2)$$

[20] The NNs that consider both CO_2 emissions and concentrations as inputs perform similarly to the other NNs for the 10-step-ahead forecasts. The MAEs (MdAEs in brackets) for the periods 1948–2007, 1968–2007 and 1992–2005 are 0.165 (0.176), 0.154 (0.143) and 0.078 (0.053), respectively. For the 20-step-ahead forecasts, the reported errors are relatively higher: 0.230 (0.206), 0.249 (0.228) and 0.169 (0.124) for the same periods.

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

Table 3
Unconditional forecasts for 10- and 20-year-ahead world annual temperature deviations.

Method	Year		Change per decade (°C)		Trend estimation per decade (°C)
	2017 ($t+10$)	2027 ($t+20$)	2017 ($t+10$)	2027 ($t+20$)	
Naïve	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

Note: The decadal trend estimation is based on fitting a linear trend on the 1- to 20-steps ahead out-of-sample forecasts of each model. The reported change per decade between 2007 and 2017 is the difference between the 10-steps ahead forecast from the last observed actuals in 2007, while the change for the second decade is the calculated difference between forecasts for 2017 and 2027.

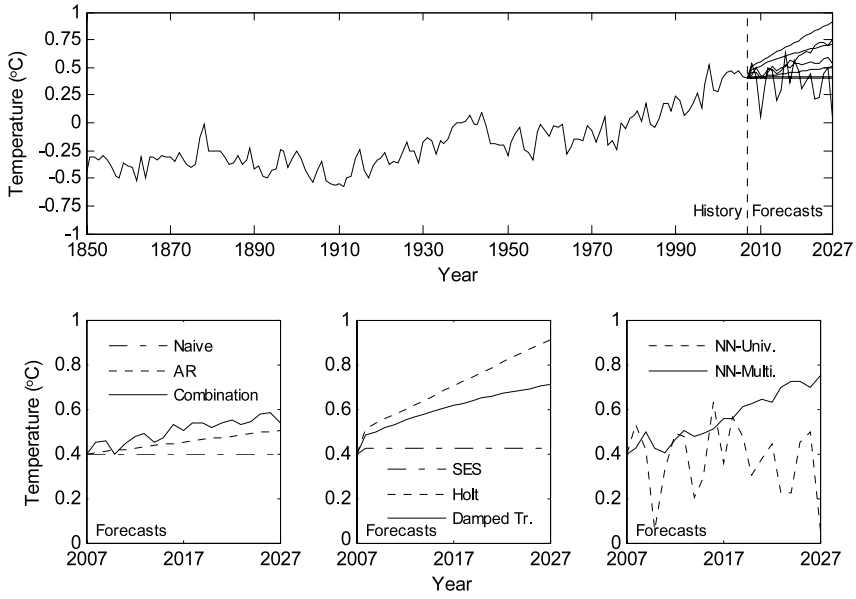

Fig. 4. 20-year-ahead world annual temperature deviation forecasts for all methods.

$$Temp_t - Temp_{t-h} = \alpha_0 + \alpha_1(ForMeth1_{t-h}(h) - Temp_{t-h}) + \alpha_2(ForMeth2_{t-h}(h) - Temp_{t-h}) + e_t, \quad (3)$$

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, $i = 1$ or 2. Eq. (1) is the standard combining approach which can also be used to test for encompassing through the test for $\alpha = 0$ (or $\alpha = 1$). Eq. (2) permits the possibility of bias and is due to Granger and Ramanathan (1984). The third equation recognizes the possibility of non-stationary data (Fang, 2003), which can be examined in either an unconstrained or a constrained form, where α_1 and α_2 must add up to 1, as in Eqs. (1) and (2). Here we examine only the constrained case, as the collinearity of the forecasts makes interpretation difficult. Note

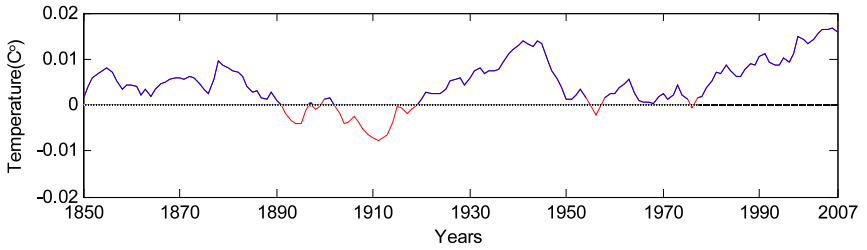

Fig. 5. Trend component estimation of the temperature deviation from the 10-year-ahead in-sample Holt forecast.

that under the constraint that α_1 and α_2 sum to 1, Eqs. (2) and (3) become identical.

In Table 4 we present the 10- and 20-year-ahead forecasts. Our focus is on establishing which methods encompass the others, if any. 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 which we consider (somewhat arbitrarily) 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 of the forecasts from the different methods.

Considering the results for Eq. (1) for both the 10- and 20-year-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 Eqs. (2) and (3), the picture is more complicated. Again, the combination of neural networks and linear models provides useful synergies; in particular, the combination of the AR and Holt methods performs very well, especially for the 10-year-ahead forecasts. For the 20-year horizon, the contribution of multivariate NNs is more apparent, providing some evidence that the effects of CO_2 become more prominent in the longer term.

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 an improvement in accuracy can be achieved through the additional information which is available from the statistical models. The proposed model is:

$$Temp_t = \alpha_0 + \sum_{i=1}^{k} a_i ForMethi_{t-10}(10) + \lambda DePresys_{t-10}(10) + e_t. \quad (4)$$

Essentially, a significant coefficient (to *ForMethi)* suggests that the GCM fails to capture the key characteristic embodied in that particular forecasting method. The combination of forecasts can be done for $1, \ldots, k$ 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 forecasts from DePreSys. 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. Since the error is stationary when $\lambda = 1$ (using an augmented Dickey-Fuller test), there is no reason to consider differences as in Eq. (3).

We present the results for the combination of each statistical method with DePreSys in Table 5. All of the combinations demonstrate improvements over the individual forecasts of DePreSys, which have a standard error of 0.103. However, only the Holt linear trend exponential smoothing forecasts seem to make any significant improvement to the accuracy, implying that the upward trend in temperature was not captured adequately in the limited period that DePreSys forecasts were available. On the other hand, the nonlinearities modelled by the equally accurate NN models do not provide significant additional new information on the 10-year-ahead forecast for that period, although the standard error of the combined forecast is improved. The constant term is insignificant, suggesting that the combined forecasts

Table 4
Forecast encompassing tests of pairs of time series models based on models (1)–(3), 10 and 20 years ahead. Standard errors are reported, and the significant forecasting methods are noted in parentheses, with A being the first, B the second and AB indicating that both are below the 5% significance level.

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 & 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 error models of the DePreSys 10-year-ahead forecasts (1992–2007). p-values are given in parentheses.

Method	Constant	Method coefficient	Standard error
Naïve	−0.149 (0.003)	+0.318 (0.206)	0.099
Single ES	−0.169 (0.003)	+0.581 (0.139)	0.096
Holt ES	**−0.260** (**0.001**)	**+0.561** (**0.014**)	**0.084**
Damped trend ES	−0.139 (0.006)	+0.243 (0.357)	0.101
AR	−0.159 (0.004)	+0.298 (0.207)	0.099
NN-univariate	−0.207 (0.006)	+0.367 (0.116)	0.095
NN-multivariate	−0.214 (0.013)	+0.338 (0.151)	0.097
Combination	−0.201 (0.004)	+0.467 (0.098)	0.094
Smith (DePreSys)	–	–	0.103

are unbiased. If the level and trend components of Holt's forecasts are considered separately, the trend exhibits a significant coefficient of +1.128, resulting in a standard error of 0.087, which is marginally worse than relying on Holt, further strengthening the argument that the DePreSys forecasts do not capture the trend exhibited in the data adequately. The level component is marginally insignificant, with a coefficient of 0.564, resulting in a reduction of the standard error to 0.093.

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

The size of the reduction in standard error is 18.4%, which is a substantial improvement in predictive accuracy, although we recognize that it is based on an in-sample fit.

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 local effects of the predicted climate change (see for example http://precis.metoffice.com/). In terms of forecast validation, they also provide a further test-bed for understanding the strengths and deficiencies of

the GCMs. Anagnostopoulos et al. (2010) and Koutsoyiannis et al. (2008) have explored this issue by evaluating various GCMs which were used in both the third and fourth IPCC assessment reports. In brief, Koutsoyiannis et al. measured the rainfall and temperature at 8 locations around the world. Six GCMs were then used to provide estimates of these quantities, and the results were 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.[22] The simulations from the GCMs are not forecasts in the same way as are Smith et al.'s carefully produced results, because they are not reinitialised through data assimilation methods at the forecast origin. Such simulations are often interpreted in much the same way as forecasts, generating arguments and policy proposals that treat the simulated values as having the same validity (or lack thereof) as forecasts. Koutsoyiannis et al. (2008) evaluate the GCM simulations at seasonal, annual and 30-year (climatic) horizons, measured via a 30-year moving average of annual temperatures. While the models capture the 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 the one- and ten-year-ahead forecasts from our time series benchmarks with the GCM forecasts.[23] The aim is to compare the 'stylised facts' in different localities with the observations, and, following on from our aggregate analysis, to see whether our time series forecasting methods add information to the local GCM-based forecasts.

The simulations were run for six (of the 8 original) localities (Albany, Athens, Colfax, Khartoum, Manaus and Matsumoto[24]) of Koutsoyiannis et al. (2008), who provided us with the local data and the GCM simulations. We use the scaled MAE, which computes the accuracy of a method as a ratio to the accuracy of the naïve random walk, and is calculated as:

$$ScaledMAE_{i,h} = \frac{\sum |Actuals_t - ForMethi_t(h)|}{\sum |Actuals_t - Actuals_{t-h}|}.$$

The closer the measure is to zero, the more accurate that method is, while if it is equal to one, the method is only as good as the random walk. We use the scaled MAE because we present the results aggregated across all six localities, and therefore the errors need to be standardised. Here, we consider a combination of the GCM-based forecasts provided by Koutsoyiannis et al. (2008). The GCM forecasts are based on a multi-model ensemble (or combination), which is calculated as an unweighted average of the different predictions from the models they describe. The spatially local results were averaged to give a measure of the overall accuracy, as is shown in Table 6. The GCM models performed substantially worse than the random walk. However, the performances of the benchmark forecasting methods used in this study were similar to or better than that of the naïve. The

Table 6
Scaled MAEs for 1- and 10-step-ahead localised temperature forecasts. The results are aggregate errors across all six localities.[25] (The data were downloaded from http://climexp.knmi.nl/. Details are available from the authors on request.)

Method	Scaled MAE	
	$t+1$	$t+10$
Naïve	1.000	1.000
Single ES	0.883	**0.901**
Holt ES	1.017	1.139
Damped trend ES	1.000	1.032
AR	0.952	0.972
NN-univariate	1.083	0.977
NN-multivariate	1.051	1.762
Combination	**0.868**	0.917
GCM	3.732	2.741

[22] The coefficient of efficiency is used in hydrology and is related to R^2 but is not so readily interpretable. It is defined as $1 - \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$ and is equal to zero if $\hat{Y}_i = Y_i$.

[23] The model setup for the benchmarks is identical to the one used to produce the global forecasts.

[24] The data ranges for the time series are 1902–2007, 1858–2007, 1870–2005, 1901–2007, 1910–2007 and 1898–2007, respectively.

[25] Different length data sets are available for each region, leading to different evaluation periods. For Albany, the evaluation period is 45 years (allowing for $45\,t+1$ and $36\,t+10$ forecasts). Similarly, the evaluation period is 89 years for Athens, 75 years for Colfax, 46 years for Khartoum, 37 years for Manaus and 49 years for Matsumoto. The accuracy over each forecast horizon for each time series is first calculated for each location and then aggregated over all localities. The data prior to the evaluation period are used for fitting the models, in the same way as for the global forecasts.

Table 7
Scaled MAEs for localised and global forecasts. The most accurate method for each horizon is in bold.

	Method	Test data for given forecast horizon		
		$t+1$ to $t+4$ 1983–2005	$t+10$ 1992–2007	$t+20$ 2002–2007
Local	Naïve	1.000	1.000	1.000
	Single ES	**0.805**	0.972	**0.890**
	Holt ES	0.905	0.960	1.080
	Damped trend ES	0.827	0.955	0.902
	AR	0.924	0.969	1.060
	NN-univariate	0.935	1.028	1.326
	NN-multivariate	0.852	0.973	1.248
	Combination	0.823	**0.886**	0.916
	GCMs	2.556	2.386	–
Global	Naïve	1.000	1.000	1.000
	Single ES	0.914	1.093	1.053
	Holt ES	**0.724**	**0.436**	0.505
	Damped trend ES	0.845	0.965	1.043
	AR	0.972	0.838	0.809
	NN-univariate	0.809	0.485	0.525
	NN-multivariate	0.845	**0.436**	**0.325**
	Combination	0.793	0.659	0.693
	Smith (DePreSys)	0.784	0.858	–

results are similar for the individual locations. This implies that the current GCM models are ill-suited to localised decadal predictions, even though they are used as inputs for policy making. The results also reinforce the need to initialise the forecasts at the forecast origin (see Mochizuki et al., 2010, for an example, although we emphasize that no 'benchmark' comparisons are made in this study of the Pacific decadal oscillation).

Using the spatially local data, we can also compare the forecasting performances of the methods relative to the random walk on a global and localised scale. This is done in Table 7, where the forecasting accuracy is shown in terms of the scaled MAE for horizons of 1–4, 10 and 20 years ahead for both the local and global forecasts. Note that the sample time periods over which the error statistics are calculated differ between Tables 6 and 7, as is described in footnote 23. It is apparent that most of the methods (with the exception of single ES and damped trend ES) can capture and model additional structure over and above that of the naïve for the global time series, resulting in significant improvements in accuracy relative to the localised GCM-based forecasts. In contrast, for the local time series, the gains from the statistical methods over the random walk are marginal, and in most cases they are unable to capture any additional structure that would result in accuracy improvements. In effect, the local variability swamps any trend, and the limited number of data points makes the 20-year-ahead results fragile. When aggregated to give world temperatures, the trend, as we have shown, becomes identifiable, which could explain the poor performance of the Holt ES and NNs. Anagnostopoulos et al. (2010) expanded the number of locations to 55 and aggregated over regions to test whether regional effects can be forecast. They reached the same conclusion as Koutsoyiannis et al. (2008): the GCMs do not produce reliable forecasts, even if aggregated to regional levels.

4. Discussion and conclusions

Decadal prediction is important both from the perspective of climate-model validation and for assessing the impact of the forecasts and the corresponding forecast errors on policy. It will also form an important part of Assessment Report 5, which is due in 2013 (Taylor, Stouffer, & Meehl, 2011; Trenberth, 2010). 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. Only a limited number of 10-year-ahead forecasts were available for evaluation (and this limitation holds across all of the still sparse decadal forecasting research). However, based on these forecasts, we have shown that the overall forecast accuracy from the DePreSys could have been improved on. More importantly, various model weaknesses were identified through the combining and encompassing analysis. In particular, adding Holt's model to the DePreSys forecasts proved of some value (decreasing the standard error by 18%). By decomposing the forecasts from Holt's model into their 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 of the DePreSys model that takes place at the forecast origin is inadequate. However, the failure to capture the local linear trend is perhaps more surprising. Other forecasting methods, and neural nets in particular, add nothing to the GCM forecasts. In essence, this suggests that the GCM captures the nonlinearities in the input-output response to emissions, but fails to capture the local trend adequately. This conclusion follows from the lack of significance of the neural net forecasts, while the linear local trend forecasts add explanatory power to the GCM forecasts. The decadal forecasting exercise appears to over-react to the forecast origin, with a smoothed value of the current system state from the exponential smoothing model providing more adequate forecasts.

Naturally, the substantive analysis we present has some serious limitations, and in particular the limited data we have gathered in relation to the DePreSys forecasts. The 10 year horizon is too short for a full decadal analysis and there are too few forecast origins included in the results from DePreSys. Because of the smoothing procedure employed by Smith et al. (2007), we have not been able to appraise their 'final' forecasts, but only their intermediate calculations. This in turn has affected our encompassing analysis, which is an in-sample analysis. In addition, there is the usual question of whether the accuracy comparisons are tainted by data snooping, whereby a comparison of a number of statistical forecasts with the GCM forecasts biases the results against the GCM. Also, we have inevitably had to focus on the only GCM of many that has been used to derive a forecast record thus far, though some others are now being used to produce such decadal data assimilated forecasts. While this limits the generality of our conclusions, we claim that none of these issues affects our overall methodological argument of the need to carry out careful forecasting exercises and corresponding forecast appraisals. Disappointingly, the latest description of the decadal modelling supporting IPCC5 (Taylor et al., 2011) suggests that, while there is to be an increased emphasis on decadal forecasting, the record being produced through data assimilation will be too short (based on 10-year-ahead forecasts produced every 5 years, starting in 1960).

The aim of this paper has been to discuss the claims relating to the validity of GCMs as a basis for medium-term decadal forecasting, and in particular, to examine the contribution that a forecasting research perspective could bring to the debate. As our analysis has shown, the DePreSys model provides 10-year-ahead forecasts that, in aggregate, could be improved by adding in statistical time series forecasts. At a more spatially 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 Anagnostopoulos et al., 2010, and Koutsoyiannis et al., 2008).

What do these comparative forecast failures imply for model validation? Within the climate modelling community 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 the model. To overcome the fact that all of the models used in the IPCC forecasting exercise have weaknesses, a combined (ensemble) forecast is produced. However, the comparative forecasting accuracy has not been given much prominence in the debate, despite its importance for both model validation and policy (Green & Armstrong, 2007; Green et al., 2009). It is surely not plausible to claim that while the decadal accuracy of GCMs is poor (relative to alternatives), their longer term performances will prove strong.

Our analysis has identified structural weaknesses in the model(s) which should point the way for climate researchers to modify either their model structure and parameterisation, or, if the focus of the modelling exercise is on decadal forecasting, the initialisation and data assimilation steps. We cannot sufficiently

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 which is relevant for policy. In carrying out the analysis reported here, we have achieved improvements in forecasting accuracy of some 18% for up to 10-year-ahead forecasts. Such improvements have major policy implications, and consequent cost savings.

Extending the horizon of decadal forecasting using a GCM to 20 years with data assimilation at the forecast origin is practical, although the computer requirements are extensive. We have also carried out a limited analysis of 20-year-ahead forecasts, though obviously without the benefit of any corresponding forecasts from a GCM. While the signal is potentially lost in the noise for the 10-year-ahead forecasts, any trend caused by emissions or other factors (see for example Pielke Sr. 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 an improved performance relative to 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 the idea is by both scientific theoretic arguments and observed improvements in predictive accuracy. The addition of the theoretically more appropriate variable, CO_2 concentration, adds little or nothing to the forecasting accuracy. However, there is no support in the evidence we present for those who reject the whole notion of global warming: the forecasts still remain inexorably upward, with forecasts which are comparable to those produced by the models used by the IPCC. The long-term climate sensitivity to a doubling of CO_2 concentration 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 the estimates arrives at a value of around 2.8, with a 95% confidence interval of 1.5–6.2 (Royer, Berner, & Park, 2007), which is compatible with the figures from the IPCC models. However, the forecasting success of a combined model composed of a GCM and a univariate time series alternative has the effect of producing a damped estimate of this sensitivity. To expand on this point, with a weighting of 0.5 on the GCM and a univariate method such as Holt, this would imply a sensitivity of just half that estimated through the GCM.

The observed short-term warming over recent decades has led most climate change sceptics to shift the terms of the political argument from questioning global warming to questioning the climate's sensitivity to CO_2 emissions. Here, we find a conflict between various of the 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 the completeness of the model as a description of the physical processes, and accordance with scientific theory and key stylised facts. In these latter cases, the GCMs perform convincingly for most in the climate modelling community. The reliance on the predictive accuracy cannot be dominant in the case of climate modelling, for the fundamental reason that the GCM models for decadal forecasting are applied to a domain which is yet to be observed. The scientific consensus is strongly supportive of the relationship between the concentration of greenhouse gases and temperature, and therefore a model needs to include such a relationship in order to be convincing outside its domain of construction. However, apparent weaknesses in the observed performances of at least one GCM have been demonstrated on shorter time scales. More importantly, the structural weaknesses in the GCM identified here suggest that a reliance on the policy implications from the general circulation models, and in particular the primary emphasis on controlling global CO_2 emissions, is misguided (a conclusion which others have reached by following a different line of argument, see Pielke Sr. et al., 2009). Whatever the success 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 of the debate 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 Dessai and Hulme (2004), Hulme and Dessai (2008) and Pielke Jr. (2003). The scientific community of global climate modellers has surely taken unnecessary risks in raising the stakes so high when depending on forecasts and models that have many weaknesses. In particular, the models may well fail in forecasting over

decades (a period which is beyond the horizons of most politicians and voters), despite their underlying explanatory strengths. A more eclectic approach to producing decadal forecasts is surely the way forward, together with a research strategy which explicitly recognizes the importance of forecasting and forecast error analysis.

Acknowledgments

We would like to thank Doug Smith for both supplying us with the data on which this paper is based and helping us to translate the language of climatologists into something closer to that of forecasters. Keith Beven and Andrew Jarvis have also helped us in the task of interpretation, not least of the comments of some highly critical reviews from the climate modelling community. A number of forecasters, environmental modellers and management scientists have helped to improve earlier drafts and have offered stimulating alternative perspectives. Critically, a referee identified an error in the data we used initially. 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.

Anagnostopoulos, G. G., Koutsoyiannis, D., Christofides, A., Efstratiadis, A., & Mamassis, N. (2010). A comparison of local and aggregated climate model outputs with observed data. *Hydrological Sciences*, *55*, 1094–1110.

Armstrong, J. S. (1985). *Long-range forecasting: from crystal ball to computer* (2nd ed.). 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.

Ascher, W. (1981). The forecasting potential of complex models. *Policy Sciences*, *13*, 247–267.

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.

Beven, K. (2009). *Environmental modelling: an uncertain future*. London: Routledge.

Bray, D., & von Storch, H. (2008). A survey of the perspectives of climate scientists concerning climate science and climate change. www.coast.gkss.de/staff/storch/pdf/CliSci2008.pdf (accessed 08/04/10).

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

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M.-F., et al. (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*, *28*, 1693–1701.

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 M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting*. Oxford: Blackwell.

Friedman, M. (1953). The methodology of positive economics. In M. Friedman (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 & 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 and 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., et al. (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 and 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.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., & Roeckner, E. (2008). Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, *453*, 84–88.

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.

Kiehl, J. T. (2007). Twentieth century climate model response and climate sensitivity. *Geophysical Research Letters*, *34*, L22710.

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.

Koutsoyiannis, D. (2010). Hess opinions "a random walk on water". *Hydrology and Earth System Sciences*, *14*, 585–601.

Koutsoyiannis, D., Efstratiadis, A., Mamassis, N., & Christofides, A. (2008). On the credibility of climate predictions. *Hydrological Sciences*, *53*, 671–684.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos, & A. Musgrave (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 S. Solomon, et al. (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. In: *Third international conference on climate change*, Washington, DC.

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, W. W., III (1972). *The limits to growth*. New York: Universe Books (in association with Potomac Associates).

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., et al. (2007). Global climate projections. In S. Solomon, et al. (Eds.), *Climate change 2007: the physical science basis. Contribution of working group 1 to the fourth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY: Cambridge U.P.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal prediction: can it be skillful? *Bulletin of the American Meteorological Society*, *90*, 1467–1485.

Mochizuki, T., Ishii, M., Kimoto, M., Chikamoto, Y., Watanabe, M., Nozawa, T., Sakamoto, T. T., et al. (2010). Pacific decadal oscillation hindcasts relevant to near-term climate prediction. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 1833–1837.

Müller, P. (2010). Constructing climate knowledge with computer models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*, 565–580.

Oreskes, N. (2003). The role of quantitative models in science. In C. D. Canham, J. J. Cole, & W. K. Lauenroth (Eds.), *Models in ecosystem science* (pp. 13–31). Princeton, NJ: Princeton U.P.

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.

Pearce, F. (2010). *The climate files*. London: Guardian Books.

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J.-M., Basile, I., et al. (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., et al. (2006). Coupled climate model appraisal: a benchmark for future studies. *Earth Observing System*, *87*(19), 185, 191–192.

Pidd, M. (2003). *Tools for thinking*. Chichester, UK: Wiley.

Pielke, R. A., Jr. (2003). The role of models in prediction for decisions. In C. D. Canham, J. J. Cole, & W. K. Lauenroth (Eds.), *Models in ecosystem science* (pp. 113–137). Princeton, NJ: Princeton U.P.

Pielke, R. A., Sr. (2005). What are climate models? What do they do? http://pielkeclimatesci.wordpress.com/2005/07/15/what-are-climate-models-what-do-they-do/ (accessed 07/07/2010).

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

Pielke, R. A., Sr. (2008). A broader view of the role of humans in the climate system. *Physics Today*, *61*, 54–55.

Pielke, R., Sr., Beven, K., Brasseur, G., Calvert, J., Chahine, M., Dickerson, R. D., Entekhabi, D., et al. (2009). Climate change: the need to consider human forcings besides greenhouse gases. *Earth Observing System Transactions, American Geophysical Union*, *90*(45), 413.

Pielke, R. A., Sr., Davey, C. A., Niyogi, D., Fall, S., Steinweg-Woods, J., Hubbard, K., Lin, X., et al. (2007). Unresolved issues with the assessment of multidecadal global land surface temperature trends. *Journal of Geophysical Research*, *112*, D24S08.

Popper, K. (2002). *The logic of scientific discovery* (2nd ed.). London: Routledge.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., et al. (2007). Climate models and their evaluation. In S. Solomon, et al. (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*, 303–311.

Royer, D. L., Berner, R. A., & Park, J. (2007). Climate sensitivity constrained by CO_2 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., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*, 403–406.

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 of London A: Mathematical, Physical and Engineering Sciences*, *365*, 2145–2161.

Stern, N. (2007). *The economics of climate change: the Stern review*. Cambridge: Cambridge University Press.

Sundberg, M. (2007). Parameterizations as boundary objects on the climate arena. *Social Studies of Science*, *37*, 473–488.

Taylor, K.E., Stouffer, R.J., & Meehl, G.A. (2011). A summary of the CMIP5 experimental design. http://www.clivar.org/organization/wgcm/references/Taylor_CMIP5.pdf (accessed February 2011).

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). Global warming and forecasts of climate change. http://blogs.nature.com/climatefeedback/2007/07/global_warming_and_forecasts_o.html (accessed 12.05.10).

Trenberth, K. (2010). *More knowledge, less certainty*. Nature Publishing Group, pp. 20–21. http://www.nature.com/climate/2010/1002/full/climate.2010.06.html.

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.

Young, P. C., & Parkinson, S. (2002). Simplicity out of complexity. In M. B. Beck (Ed.), *Environmental foresight and models: a manifesto* (pp. 251–294). Oxford: Elsevier.

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.

Robert Fildes is Distinguished Professor of Management Science in the School of Management, Lancaster University and Director of the Lancaster Centre for Forecasting. He was co-founder in 1981 of the *Journal of Forecasting* and in l985 of the *International Journal of Forecasting*. For ten years from l988 he was Editor-in-Chief of the IJF. He was president of the *International Instititute of Forecasters* between 2000 and 2004. His current research interests are concerned with the comparative evaluation of different forecasting methods, the implementation of improved forecasting procedures in organizations and the design of forecasting systems. His interest in climate modelling arose from the realization that the forecasting community has made little contribution to the important debate about global warming.

Nikolaos Kourentzes is a post-doctoral research assistant in Management Science at Lancaster University Management School. He received his Ptychion degree from Athens University of Economics and Business (AUEB), an M.Sc. in Management Science and a Ph.D. from Lancaster University Management School. His research focus is on time series prediction, with a particular emphasis on neural networks, input variable selection and high frequency data.