TECHNICAL MEMORANDUM

518

Using NWP to assess climate models

M. J. Rodwell and T. N. Palmer

Research Department

Based on a paper published in Quart.J. Roy.Meteor.Soc.

March 2007

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

SUMMARY

Estimates of climate change remain uncertain - hampering strategic decision making in many sectors. In large part this uncertainty arises from uncertainty in the computational representation of known physical processes. This model component of climate change uncertainty is increasingly being assessed using perturbed model experiments. Some such model perturbations have, for example, led to headline global warming estimates of as much as 12°C. These experiments consider many differently perturbed versions of a given base model and assess the likelihood of each perturbed model's climate prediction based on how well it simulates present-day climate. In these experiments, the computational cost of the model assessment is extremely high unless one assumes that the climate anomalies associated with different model perturbations can be combined linearly. Here we demonstrate a different method that harnesses the power of the data assimilation system to directly assess the perturbed physics of a model. Data assimilation involves the incorporation of daily observations to produce initial conditions (analyses) for numerical weather prediction (NWP). The method used here quantifies systematic initial tendencies in the first few timesteps of a model forecast. After suitable temporal averaging, these initial tendencies imply systematic imbalances in the physical processes associated with model error. We show how these tendencies can be used to produce probability weightings for each model that could be used in the construction of p.d.f.s of climate change. The approach typically costs 5% of the cost of a 100-year coupled model simulation that might otherwise be used to assess the simulation of present-day climate. Importantly, since the approach is amenable to linear analysis, it could further reduce the cost of model assessment by several orders of magnitude: making the exercise computationally feasible. The initial tendency approach is only able to assess "fast physics" perturbations, i.e. perturbations that have an impact on weather forecasts as well as climate. However, recent publications suggest that the majority of present model parameter uncertainty is associated with fast physics. If such a test were adopted, the assessment of the ability to simulate present-day climate would then only be required for models that "pass" the fast physics test. The study highlights the advantages of a more seamless approach to forecasting that combines numerical weather prediction, climate forecasting and all scales in-between.

KEYWORDS: Climate change, Perturbed model, Data assimilation, Initial tendency, Analysis increment, Model imbalance, Linearity

1. Introduction

Anthropogenic climate change is one of the biggest challenges faced by the world today. Decisions about global emissions of greenhouse gases are highly reliant on the perceived accuracy of climate forecasts. However the uncertainty in the prediction of global -mean surface temperature has remained little changed through the course of the first three Intergovernmental Panel on Climate Change (IPCC) assessment reports. The first assessment report (FAR) estimated a range for equilibrium warming due to a doubling of carbon-dioxide (CO_2) of 1.5-4.5K (Mitchell et al., 1990). The FAR was based on experiments with atmospheric general circulation models (AGCMs) coupled to rather simple mixed-layer ocean models. Much of the climate uncertainty was attributed to uncertainties in the representation of cloud radiative feedbacks. The second assessment report (SAR) estimated a range of 1.0-3.5K (Kattenberg et al., 1996) although this was based on transient CO_2 experiments rather than equilibrium experiments and the models incorporated a dynamic ocean component. This range represented continued large model uncertainties as well as uncertainties in projected emissions of trace gases and aerosols. The general reduction in climate sensitivity relative to the FAR was attributed to the inclusion of the direct and indirect effects of sulphate aerosol. The third assessment report (TAR) estimated a range for global warming of 1.4-5.8K (Cubasch et al., 2001). Again this range was associated with model and emission uncertainty. The higher maximum value relative to the SAR was mainly due to lower projected sulphur-dioxide (SO2) emissions (Wigley and Raper, 2002).

Computational constraints make it difficult to reduce model uncertainty by substantially increasing model resolution. Hence forecasters must rely on improving the representation (parametrization) of sub-grid scale processes such as atmospheric convection or oceanic eddy heat fluxes. One would hope that there was a tendency, over the course of the first three assessment reports, for the representation of processes within climate models to begin to converge to the physical processes of the real world. However, it could be argued that there was also a possibility for climate models to converge for another reason: as modellers consciously or unconsciously attempted to make their own models come into better agreement with what may have been perceived to have been the better models of the day. There is a risk that the range of global warming estimates may be underestimated by this process.

There are other potential sources of uncertainty that may be omitted in the above climate change estimates. For example, there is uncertainty in the degree to which the large-scale flow can constrain sub-grid scale processes (Buizza et al. 1999, Palmer 2001, Palmer et al. 2005). Also, uncertainty in the initial conditions of features such as the oceans and ice sheets may not have been fully represented. The uncertainty associated with a lack of knowledge of what processes have been left-out of models may be partially estimated by assessing a model's ability to simulate recent climate variations (Tett et al.; 1999, Andronova and Schlesinger; 2001). However, the applicability of such an approach depends on whether the short observational record is sufficient to capture low-frequency variations of the oceans and ice-sheets and whether processes combine linearly and behave similarly in a warmer climate.

In all three reports, the word "range" generally refers to the range of all possible climate sensitivities of a somewhat ad-hoc collection of climate models. Ideally, for each anthropogenic emission scenario, what is required is a probability distribution function (p.d.f.) of climate change that reflects uncertainties associated with deterministic chaos, non-deterministic sub-grid scale variability, and present-day uncertainties in our knowledge of the science. The importance to policy makers of improved p.d.f.s of climate change was underlined in the TAR, which called for the development of better methods of assessing uncertainty (IPCC Summary for Policymakers, Houghton et al., 2001).

Murphy et al. (2004) investigated a "perturbed model ensemble" methodology that represents a first step towards a more systematic approach to assessing the present p.d.f. of climate change. They asked experts to estimate the range of uncertainty associated with a set of tuneable model parameters. They perturbed these parameters (one by one) in their model (AGCM coupled to a mixed-layer ocean model) to produce a 53-member perturbed model ensemble. Present-day climate and $2xCO_2$ simulations were made to deduce each model's equilibrium "climate sensitivity" to a doubling of CO_2. By assuming that the climates of differently perturbed models combine linearly, they claimed an effective ensemble size of $4x10^6$ model versions. By "combine linearly" we mean that, relative to some control climate, the anomalous climate of a model with perturbation "A" plus the anomalous climate of a model with perturbation "B" is equal to the anomalous climate of a model with both perturbations A and B. The anomalous climate of each of the $4x10^6$ model versions was calculated from the 53 pairs of simulations assuming this linearity. A weighted p.d.f. was constructed by weighting the climate sensitivities of the $4x10^6$ member ensemble by the accuracy of their (linearly combined) present-day climates. An unweighted p.d.f. was constructed by assuming that all 4x106 model versions were equally likely. The authors found that weighting led to a shift in the peak of the p.d.f. of climate sensitivity from 2.5K to over 3.2K. There was also a modest narrowing of the p.d.f. with the 90% probability range being left at 2.4-5.4K. Clearly weighting is an important issue in the construction of p.d.f.s of climate change.

A present-day climate assessment involves long simulations of the full climate model, which ideally should incorporate atmosphere, dynamic-ocean, ice, vegetation, chemistry, etc., components. These simulations are computationally costly and, without Murphy et al's linearity assumption, this cost will limit the number of model versions that can be assessed. Whether the climates of perturbed models do combine linearly is clearly a vital question for future research.

Stainforth et al. (2005) used the "climateprediction.net" method of harnessing the idle processing capacity of personal computers throughout the world to make a similar study with 2017 perturbed versions of the Murphy et al. (2004) model. Importantly, the increased computing power allowed them to drop the climate linearity assumption. Although it may be a long while before personal computers are individually powerful enough to run high-resolution, coupled climate models, the distributed technique is a useful research tool and the public educational aspect of "climateprediction.net" should not be underestimated.

Another potentially more efficient approach to the systematic assessment of model uncertainty was investigated by Annan et al., (2005b). They used an Ensemble Kalman Filter data assimilation scheme where some model parameter values were included as part of the model state space and where the cost function involved a test of the model's simulation of present-day climate. In theory their method can produce a joint distribution for the parameter values. For a simplified atmospheric model they were able to reproduce 3 or 4 out of 5 known parameter values in "identical twin" experiments where the model is assumed to be perfect (except for the few unknown parameter values). Further work is required to determine the efficiency of this approach when applied to more complex, more non-linear and less perfect models, with more tuneable parameters.

A good simulation of present-day climate is clearly an essential prerequisite for establishing faith in a model's prediction of climate change. However, it would be very useful to develop additional complementary tests that may be more efficient or that may be able to assess more directly the underlying physics within a model. At present, much of the perceived uncertainty in model physics involves "fast" processes that are also important in numerical weather prediction (NWP). Table 1, for example, shows a representative sample of the physics parameter uncertainties that Murphy et al., (2004) assessed. Around 80% of the model parameters that they perturbed (the first 8 out of 10 in Table 1) are associated with "fast physics". Hence this raises the possibility that NWP techniques could be used to assess climate models.

Phillips et al (2004) initialised a climate model with the European Centre for Medium-range Weather Forecasts (ECMWF) re-analysis dataset "ERA-40" (Uppala et al, 2005) and diagnosed the mean ("systematic") error after the first 5 days of a forecast. They argued that since the initial conditions are close to the "truth", this systematic error must be attributable to parametrization deficiencies. Tests with a proposed change to the triggering of convection were shown to reduce this short-range error. It was argued that a reduction in systematic 5-day forecast error implies that a climate model is more physically realistic than its predecessor.

Here, we take model physics assessment to its ultimate limit by considering systematic tendency errors over the first few forecast timesteps. At this timescale, we are able to demonstrate that the assessment is linear enough to estimate the impact of a set of perturbations to different parametrizations as the linear sum of their individual impacts. In this way, it becomes computationally feasible to incorporate a probability of the realism of each model's fast physics into the p.d.f. of climate change.

Table 1 Some model parameters perturbed by murphy et al. (2004)

Parameter	Physical Process	Values Used		
		Low	Middle	High
Droplet to rain conversion rate (s^{-1})	Cloud	$0.5x10^{-4}$	$1.0x10^{-4}$	$4.0x10^{-4}$
Relative humidity for cloud formation	Cloud	0.6	0.7	0.9
Cloud fraction at saturation (free trop.)	Cloud	0.5	0.7	0.8
Entrainment rate coefficient	Convection	0.6	3.0	9.0
Time-scale for destruction of CAPE (h)	Convection	1.0	2.0	4.0
Effective radius of ice particles (μm)	Radiation	25	30	40
Diffusion e-folding time (h)	Dynamics	6	12	24
Roughness length parameter (Charnock)	Boundary	0.012	0.016	0.020
Stomatal conductance dependent on CO_2	Land	Off	-	On
Ocean-to-ice heat diffusion coefficient (m^2s^{-1})	Sea Ice	$2.5x10^{-5}$	$1.0x10^{-4}$	$3.8x10^{-4}$

A representative list of the model parameters perturbed by Murphy et al. (2004) together with the physical process they are associated with and the perturbed values used.

As forecast lead-time increases, anomalies in an atmospheric state vector become increasingly constrained by the need for consistency. This consistency is imposed by, for example, synoptic organisation and planetary teleconnections. By concentrating on the first few forecast timesteps, these dynamical constraints are minimised, and allow the space of initial tendencies to have a very large dimension (perhaps as large as the number of grid-points or spherical harmonics in the model for example). With such a large dimension, it becomes very difficult to find two distinct model errors that can compensate each other. In this way the hope is that the initial tendency approach can be a very discriminating way of assessing model physics.

Another difference with Phillips et al. (2004) is that initial states are not associated with some fixed "control" model, rather data is assimilated into each perturbed model in turn. We consider this to be essential aspect of our study because the analysed state used for forecast initiation and verification can be sensitive to the model used in the data assimilation process; we don't want to somehow bias our results to the choice of model used for data assimilation. Arguably, this is a more critical requirement for our study based on short-range tendencies, than in Phillips et al., where errors have, by day 5, probably grown to be much larger than the differences between analyses produced using different assimilating models.

The structure of the paper is as follows. In section 2 we motivate and explain developments of the initial tendency methodology. In section 3 we discuss the perturbed models that we will analyse in detail. In section 4 we discuss conventional measures of NWP forecast verification and suggest that these are not perfectly suited to model physics assessment. Section 5 highlights, for a particular region, the strong connection between physical changes to a model and changes in the initial forecast tendencies. In section 6, we define a single global score of a model's physics based on initial tendencies. The linearity of the initial tendency approach is investigated in section 7 and the computational cost of the approach is given in section 8. Section 9 discusses the practical implementation of an initial tendency test in order to produce p.d.f.s of climate sensitivity. In section 10 we demonstrate that a model that 'fails' our initial tendency test actually passes a conventional present-day climate test - further emphasizing the need for additional methods such as ours that specifically assess aspects of model physics. A discussion and conclusions are given in section 11. In particular we highlight the need for a more 'seamless' approach to weather and climate forecasting.

2. Methodology

2.1 Motivation

Numerical weather prediction involves the use of highly complex forecast models. As these models improve, we need continually to look for ever more precise ways of identifying remaining deficiencies and of comparing model versions. In this subsection, we give some motivation for one such method based on initial tendencies.

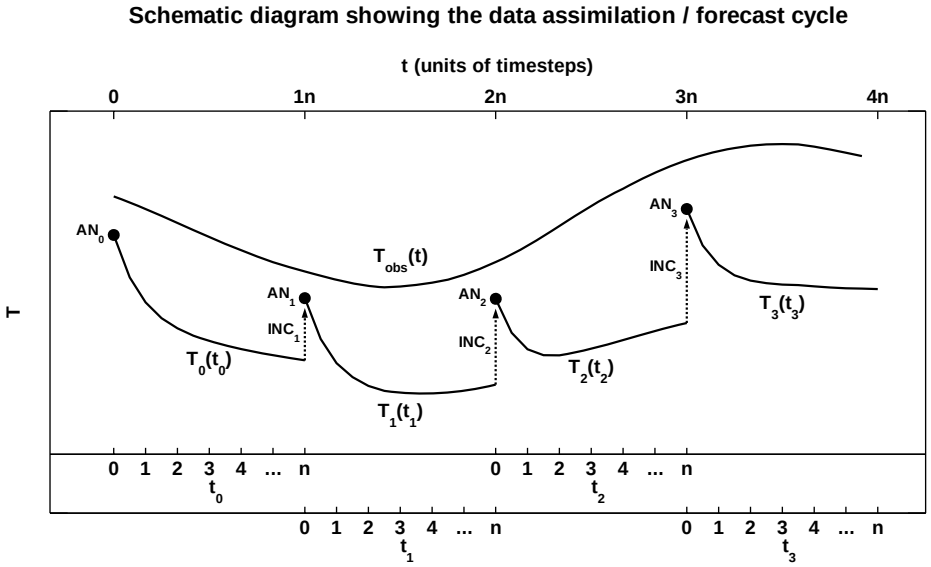

Figure 1 Schematic diagram showing the data assimilation and forecast integration aspects of numerical weather prediction. $T_{obs}(t)$ represents an observed timeseries (e.g. of temperature at some specified location). For each i , $T_i(t_i)$ represents the model forecast initiated from analysis AN_i. *For the purposes of explaining our methodology, the role of systematic forecast error (in this case a cooling) has been emphasised over random error. See the main text for further explanation.*

NWP generally involves two main processes: data assimilation and forecast model integration. Fig. 1 shows a schematic diagram of these processes. The observation curve, T_{OBS}, could represent the observed temperature at some fixed location as a function of time, t. (Here time is measured in units of forecast model timesteps). Observations are generally inaccurate, unbalanced and incomplete and so it is neither sensible nor possible to initiate a forecast model directly from the observations. Instead, a model "first guess" from a previous forecast $T_0(t_0)$ is combined with the global observations over some time-window to produce a "best estimate" (or "Analysis", *AN*) of the present state of the atmosphere. This process is known as data assimilation. The implications of this study are not thought to depend on the precise details of how the data assimilation is done and so further information on the ECMWF assimilation system (Rabier et al., 2000) is not given here. The forecast $T_1(t_1)$ is initiated from the analysis AN_1. In turn, $T_1(t_1)$ is used as the first guess for the next analysis AN_2, etc. Each first guess forecast could, of course, be extended to produce a longer-range weather forecast. Note that for the purposes of explaining our methodology, we may have emphasised in Fig. 1 the systematic component of forecast error (i.e. a cooling if *T* represents temperature) over that of random error. Fig. 2 will show the relative magnitudes of tendencies associated with all processes using actual forecast data.

An important measure of the quality of the forecast system is the "analysis increment", *INC*. The analysis increment is the increment applied to the first guess state to get to the new analysis state. If there are *n*

forecast timesteps between the start of the first guess forecast and the new analysis then the increment for AN_1 can be written as:

$$\begin{aligned} INC_1 &= AN_1 - T_0(n) \\ &= T_1(0) - T_0(n) \quad . \end{aligned}$$

More generally:

$$INC_i = T_i(0) - T_{i-1}(n) \quad .$$

If we had a perfect model and perfect observations then this should result in an analysis increment of zero. (Because the forecast curves would lie on top of the observation curve). In reality, observations do have errors but bias-correction prior to data assimilation should result in these errors being random with zero mean. Hence even for non-perfect observations, a perfect model should result in the *mean* analysis increment (averaged over many data assimilation / forecast cycles) being close to zero. If the mean increment is not zero, then this is indicative of "model spin-up". Model spin-up is associated with errors in the model's representation of the physics of the atmosphere (or the lack of representation of a physical process). Hence, subject to certain conditions (see below), it can be argued that the smaller the mean analysis increment, the closer the first guess is to the observations and the better is the forecast model.

The mean analysis increment over m data assimilation / forecast cycles can be written as:

$$\begin{aligned} \frac{1}{\mathrm{m}}\sum_{i=1}^{m} INC_i &= \frac{1}{m}\sum_{i=1}^{m}\left(T_i(0) - T_{i-1}(n)\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left(T_i(n) - T_i(0)\right) + \frac{1}{m}\left(T_m(n) - T_0(n)\right) \quad . \end{aligned}$$

For large m , the end-point term $\frac{1}{m}\left(T_m(n) - T_0(n)\right)$ becomes vanishingly small and so we can write:

$$\begin{aligned} \frac{1}{\mathrm{m}}\sum_{i=1}^{m} INC_i &\approx -\frac{1}{m}\sum_{i=1}^{m}\left(T_i(n) - T_i(0)\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(T_i(j) - T_i(j-1)\right) \\ &= -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\dot{T}_i(j) \\ &\equiv -\bar{\dot{T}} \quad , \end{aligned}$$

where an over-dot indicates a time-derivative (in units of K per timestep, for example) and an over-bar represents the mean over all assimilation/forecast cycles as indicated. Hence, for large m, the mean analysis increment is almost precisely the negative of the "systematic forecast tendency", $\bar{\dot{T}}$. This link can be readily seen in the schematic diagram where the model is consistently moving to a cooler state and the analysis increment adjusts the state back to a warmer one. Hence systematic forecast tendencies can also be used to gauge the quality of a forecast model.

Klinker and Sardeshmukh (1992), hereafter KS92, attempted to take the systematic tendency approach further by considering just the tendencies in the very first forecast timestep:

$$\frac{1}{m}\sum_{i=1}^{m}\dot{T}_i(0) \quad .$$

Implicit in their methodology was an accounting-for of the mean diurnal cycle by using a 6-hr data assimilation / forecast cycling period; see below. Justification for considering just the first timestep this is that if the systematic tendency in the first timestep is small (or even zero) then so should be the systematic tendencies in the subsequent timesteps. An advantage of considering only the first timestep is that the initial tendencies are evaluated at a state as close to the true state as possible and so errors in $\dot{T}_i(0)$ should be more directly associated with deficiencies in the model processes and less associated with the application of these processes to erroneous atmospheric states.

We can write $\dot{T}_i(0)$ as the sum of the individual tendencies associated with the model's representations of the dynamics (dyn), radiation (rad), convection (con), large-scale precipitation (lsp), and all other processes as:

$$\dot{T}_i(0) = \dot{T}_i^{dyn}(0) + \dot{T}_i^{rad}(0) + \dot{T}_i^{con}(0) + \dot{T}_i^{lsp}(0) + \ldots \quad .$$

The "dynamical tendency" generally refers to the tendencies that are resolved on the model grid. The parametrized tendencies such as the "convective tendency" are often all referred to as tendencies associated with model "physics". Here we will refer to the dynamics, convection etc as "physical processes", or simply "processes", with little distinction between whether they are resolved or parametrized. Other numerical adjustments can also occur within a model timestep. The tendencies associated with these adjustments should also be quantified but will probably not have a strong effect or change any conclusions (they do not in our studies). It can be seen that the "total systematic initial tendency", $\bar{\dot{T}}_i(0)$, is the (possibly small) residual of a balance between (large) physical processes. If $\bar{\dot{T}}_i(0)$ is non-zero then, in a time-mean sense, these processes are "out-of-balance" at the analysis state. Since each individual process is initially acting on a state close to the truth, errors have not had time to interact or propagate and so it may be possible to identify which model process(es) lead to the erroneous total systematic tendency. For example, if the cooling seen in the schematic diagram only occurs in convective regions (i.e. if $\bar{\dot{T}}_i^{con}(0)$ has a similar spatial pattern to $\bar{\dot{T}}_i(0)$) then this may indicate erroneously weak convective latent heat release for a given thermal and humidity profile. Clearly this initial tendency approach could potentially form the basis of an automated search for model error. By applying this search to the momentum budget, KS92 identified problems with the ECMWF gravity wave drag scheme. Experience shows that the search for thermodynamic model errors with this approach is more difficult; possibly because there is more scope for multiple independent thermodynamic errors to complicate a simple interpretation of the total initial tendency. In this paper, we do not attempt to identify model errors. Instead, we focus on the assessment of model physics *after* a model change has been made. This methodology has been very successful in demonstrating conclusively that a particular change to the model was physically justified (Rodwell and Jung, 2007).

In the above discussion, it was stated that the smaller the mean analysis increment (or almost equivalently the smaller the systematic initial tendency), the better the representation of the dynamics and physics within the model. There are several important conditions and comments to attach to this statement.

1. If no observations were used in the assimilation system then the analysis would be equal to the first guess and so, for a fully spun-up model, the mean analysis increment and systematic tendency would both be zero; even for an *imperfect* model. Hence it is clear that the observations are crucial to the success of the methodology.

2. It is possible that there could be a "compensation of errors" within a model, so that two or more rather large physics errors produce tendency errors that cancel each other out. If this were to happen

then the systematic initial tendency could be small despite the physics being poorly represented. However, since we will be interested in initial tendencies world-wide, this cancellation would have to occur in a very high dimensional space; at every grid-point and every model level. This is highly unlikely because different physical processes (and therefore their tendency errors) are likely to be dominant in different regions and/or at different altitudes. Put another way, there would be no point for the inclusion in a model of a new process whose tendencies can be represented by the processes already in the model. Hence, from the very way models are developed, complete cancellation is precluded.

3. Another possibility for compensation is if a single model error leads to inaccuracies in the analysis which, when other physical processes are applied to this analysis, lead to compensating tendencies. However, since the observations constrain the analysis, complete compensation via this route is also unlikely to occur. This constraint by the observations is fundamentally the reason why it is beneficial to use NWP to assess the fast physics of climate models. Over each assimilation cycle, the observations continually draw the atmospheric state towards the truth and thus shift an inaccurate model to a state of imbalance.

The above points highlight the importance of the observations and the limits to which the initial tendency approach can be used to assess model physics. It is clearly worth experimenting with the application of the methodology to the assessment of fast physics (as done here) but these limits are one reason why other tests of climate models will remain essential.

2.2 Accounting for the diurnal cycle

There is a potential problem from only considering the first forecast timestep when estimating systematic tendencies. While averaging over many assimilation/forecast cycles should reduce the impact on the estimated systematic tendency associated with synoptic and longer-timescale variability, it may not remove the tendency associated with the mean diurnal cycle. For example, if the cycling period, timesteps, is equal to one day with the analysis valid at, say, 6am local time, then the mean temperature tendency in the first timestep will reflect legitimate warming at sunrise as well as systematic tendency errors. KS92 used a 6-hour cycling period and averaging over the mean tendencies at 00, 06, 12 and 18 UTC should reduce the impact of the diurnal cycle. A sinusoidal diurnal cycle, for example, would be perfectly removed regardless of local time although more asymmetric diurnal cycles could cause problems.

Using just the first timestep, KS92 were successful at identifying possible reasons for systematic momentum tendencies. Here, however, we will also be interested in thermal tendencies whose diurnal cycles may have larger amplitudes and be more asymmetric. In addition, 14 years have elapsed since the study of KS92 and, while forecast system improvements have reduced the size of systematic tendency errors, the magnitude of potential errors associated with a coarse accounting of the diurnal cycle will remain unchanged. Before using systematic initial tendencies to assess perturbed models, we first investigate the impact of the diurnal cycle.

While systematic *initial* tendencies may combine a diurnal cycle effect with model spin-up, systematic tendencies later in the forecast (after the model has spun-up) will only include the diurnal effect. Because of this, we look at tendencies in the ECMWF model (see later for more details on the model) at a lead-time of around 5 days (D+5) to diagnose the diurnal effect. We focus on the Amazon/Brazil region (300°E-320°E, 20°S-0°N) in southern summer because the South American monsoon has been historically problematic for the ECMWF model. Forecasts are started every 6 hours from 00 UTC on 27 December 2004 until 18 UTC on 26 January 2005 (31 days x 4 forecasts per day = 124 forecasts). The forecast timestep is ½ hour. Fig. 2a

shows temperature tendencies on a model level at approximately 353hPa (T353) averaged over the Amazon/Brazil region. The solid curve shows a concatenation of the tendencies between D+5 (actually timestep 241) and D+5¼ (actually timestep 252) for the first 20 forecasts. A strong and asymmetric diurnal cycle in tendencies is clearly visible with intense heating over a 6 hour window in the morning and more gradual cooling throughout the rest of the day. The circles in Fig. 2a show the diurnal average temperature tendency averaged over all 48 ½-hour timesteps (12 timesteps x 4 forecasts). These diurnal averages appear to be close to zero. The average over all 31 days is slightly negative (-0.06±0.05 $Kday^{-1}$). The range of possible values (-0.11 to -0.01 $Kday^{-1}$) represents the 70% confidence interval based on the Student t-distribution taking autocorrelation into account, see von Storch and Zwiers (2001). This confidence interval is indicated by the right-hand bar in Fig. 2b. Hence when all 48 timesteps are taken into account there is, if anything, a very slight systematic cooling, possibly associated with the later stages of model spin-up and/or the gradual annual cycle.

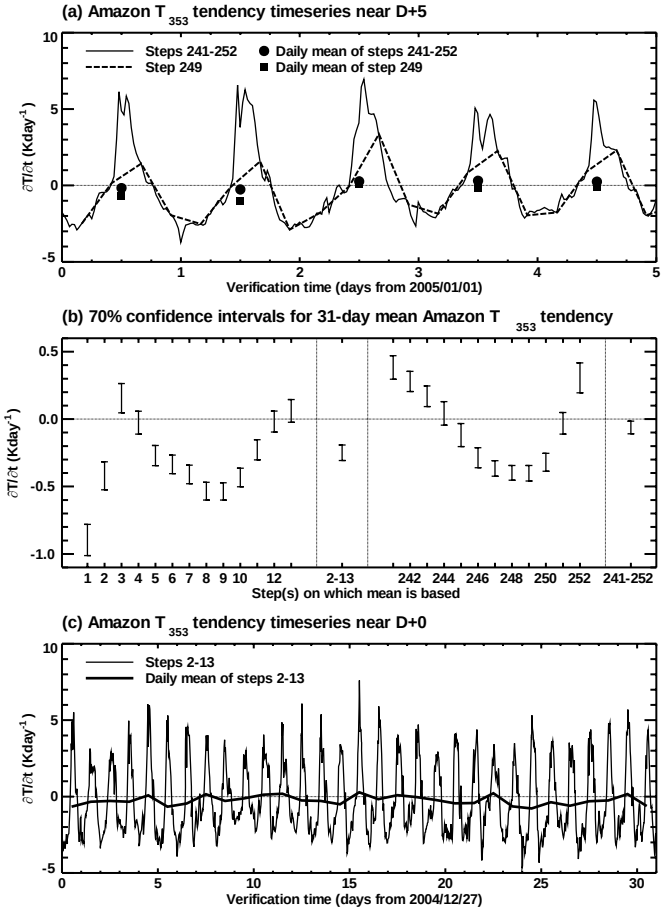

Figure 2(a) The thin solid curve shows a concatenation of Amazon/Brazil temperature tendency timeseries at approximately 353 hPa ($\dot{T}_{353}$) based on timesteps 241 to 252 of forecasts initiated every 6 hours from 00 UTC on 2004/12/27 (data from 20 forecasts are shown). The dashed line shows the same timeseries sub-sampled at forecast timestep 249 only. Filled circles (squares) how the daily-mean values of the solid (dashed) curves. (b) 70% confidence intervals for the daily-mean tendency based on the forecast timestep(s) indicated on the x axis and using all 124 forecasts initiated every 6 hours between 00 UTC on 2004/12/27 and 18 UTC on 2005/01/26. (c) The thin solid curve shows a concatenation of $\dot{T}_{353}$ timeseries based on timesteps 2 to 13 of all 124 forecasts. The thick solid curve shows the daily-mean of the thin curve. The Amazon/Brazil region is defined as 300°E-320°E, 20°S-0°N.

If just one timestep is used from each forecast then the diurnal cycle is sampled 4 times a day. Choosing timestep 249, the diurnal cycle looks like the dashed curve in Fig. 2a. In this case, the strong morning heating is not well sampled and this makes the estimated daily cooling too strong (squares in Fig. 2a). Fig. 2b also shows the diurnal-mean tendencies estimated using each of the timesteps 241 to 252 individually. The estimated tendencies show variations that reflect changes in the sampling of the diurnal cycle. The question is, are these variations small enough to be able to claim that the initial tendencies are predominantly due to model error and not due to errors in diurnal sampling? Fig. 2b also shows mean tendencies based on the individual timesteps 1 to 13. It is clear that these are comparable in magnitude to the tendencies based on the

single timesteps 241 to 252. We must conclude, therefore, that it is no longer sufficient to sample the diurnal cycle using just four points.

Another problem is highlighted in Fig. 2b: the first timestep is very different from subsequent timesteps. This is a consequence of structural differences in the first timestep that are essential to allow the model to "cold start" from the analysis. For example, the semi-Lagrangian advection scheme in the ECMWF model ordinarily requires the model tendencies from the previous timestep. A different approach has to be taken in the first timestep as these tendencies are not available. This and other differences mean that the model in the first timestep is effectively a somewhat different model to that of subsequent timesteps.

Because we really wish to diagnose the error associated with the model used after the first timestep (i.e. the model that is used in the climate forecast) and because a single timestep is not sufficient to account for the diurnal cycle, we average over timesteps 2-13. The mean temperature tendency for T_{353} over the Amazon/Brazil region based on steps 2-13 is -0.25±0.06 $Kday^{-1}$ (also indicated in Fig. 2b). Since this is large in magnitude compared to the -0.06±0.05 $Kday^{-1}$ cooling seen between timesteps 241-252, we can be reasonably sure that this represents a cooling due to systematic model error.

To further emphasise the relative magnitudes of the typical signal we are trying to isolate and those of the other variations in the system, Fig. 2c shows the concatenated Amazon/Brazil T_{353} tendencies based on timesteps 2-13 from all 31x4 forecasts (thin) and the daily-averaged tendencies (thick). The negative mean tendency associated with model error (-0.25±0.06 $Kday^{-1}$) is reflected in the fact that the thick line remains below zero for much of the month. It is clear, however, that we are looking (in the un-perturbed model at least) for small mean tendencies in relation to diurnal and, to a lesser extent, synoptic variability. Because of this, we will make use of confidence intervals to demonstrate that a signal has (or has not) been identified. Although Fig. 2c suggests that the trend due to the annual cycle is very small (for this region and time of year), any such trend would also be reflected in the confidence intervals.

2.3 Summary of present methodology

We define the "systematic initial tendency" as the mean forecast tendency averaged over the twelve ½-hour timesteps 2 to 13 for forecasts started every 6-hours. Using timesteps 2-13, our systematic initial tendencies is very similar to the mean analysis increment. Hence, while bearing in mind comments (1) to (4) in section 2a above, the smaller the systematic initial tendency, the better the representation of dynamics and physics within the model. The hope is that ~6 hours is still short enough for there to be minimal interaction between the processes represented in the model. This will be seen as a crucial aspect of the methodology; making the estimation of climate change sensitivity to model error a computationally manageable problem.

3. The models and integrations

3.1 Control model

The base model used here is the recently operational atmospheric model (version 29R1) from ECMWF. This control model is referred to as the "CONTROL".

3.2 Perturbed models

Stainforth et al. (2005) made their present-day climate assessment of each model by calculating the average of the spatial root-mean-square errors in the simulation of 8-year-mean surface temperature, sea-level pressure, precipitation and surface sensible and latent heat fluxes (relative to those of their base model). The circles (all colours) in Fig. 3 show present-day climate error versus the climate sensitivity (global-mean

surface temperature change arising from a doubling of CO_2) for each perturbed model. By comparing with the present-day climate errors of other coupled models, Stainforth et al. (2005) concluded that none of the perturbed models could be rejected. Hence, climate sensitivities of as much as 11K are possible within their framework. The blue circles in Fig. 3 correspond to all the perturbed models where, amongst other possible perturbations, the "entrainment rate coefficient" (which sets the rate at which moisture is turbulently exchanged between the convective plume and its environment) has been reduced by a factor 5 (see Table 1). It can be seen that the highest climate sensitivities come from these perturbed models. The climate sensitivities for the Murphy et al. (2004) models also show a similar relationship between climate sensitivity and the entrainment rate coefficient (personal communication James Murphy). If one could reject the low entrainment perturbation of Stainforth et al. (2005), then the uncertainty in the climate sensitivity associated with model parameter uncertainty would be greatly reduced. With this motivation in mind, we demonstrate our initial tendency method for the assessment of model physics using similar perturbations to the entrainment coefficient. It should be stressed, however, that we are using a different base model to that of Stainforth et al. (2005) and our aim is to demonstrate a methodology rather than to constrain their climate sensitivity range. As with Stainforth et al. (2005), we divide the entrainment coefficient by 5 for our low-value experiments (termed "ENTRAIN/5") and multiply by 3 for our high-value experiments (termed "ENTRAINx3").

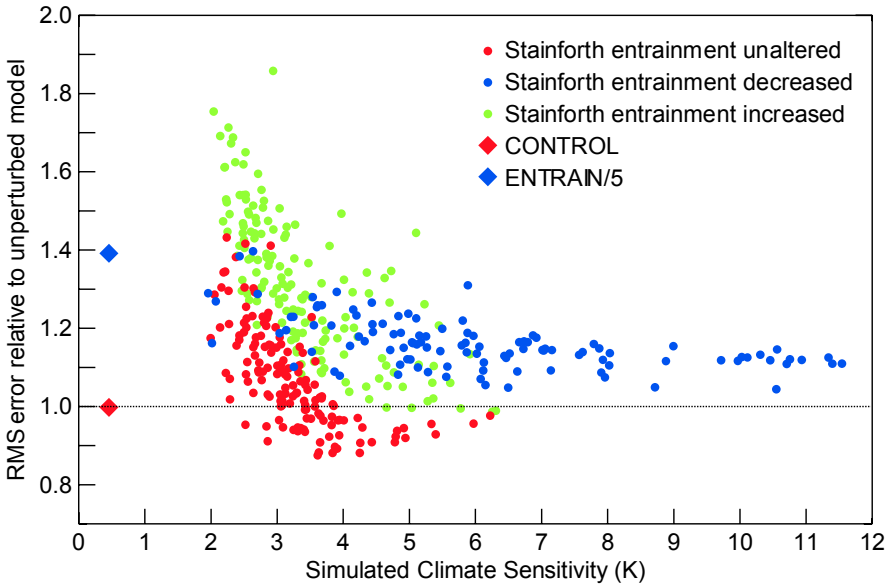

Figure 3 The root-mean-square error of various climate models plotted against their global-mean surface temperature sensitivity to a doubling of CO_2. Circles represent the atmosphere/mixed-layer ocean models used by Stainforth et al. (2005). Blue circles represent perturbed models where the convective entrainment coefficient was reduced by a factor 5. Green circles represent perturbed models where the entrainment coefficient was increased by a factor 3. The red circles represent perturbed models where the entrainment coefficient was not altered. Diamonds represent two of the atmospheric models investigated in this study. Red diamond: the CONTROL model. Blue diamond: the ENTRAIN/5 model where the convective entrainment coefficient was reduced by a factor 5. The plotted error is actually the mean of the root-mean-square errors (RMSEs) for a number of different parameters, each normalised by the control model RMSEs for each parameter. See Stainforth et al. (2005) for further details.

An additional model version (termed "CLOUD") is tested alongside these perturbed models. CLOUD incorporates a structural change to the cloud scheme that is being tested at ECMWF for incorporation into the operational model (Tompkins et al. 2006). CLOUD is meant to represent here a "typical" modification

that may be made during NWP development. (It is possible to assess structural as well as parameter changes with the initial tendency methodology).

Further perturbed models are discussed in section 7 where they are used to assess the linearity of the initial tendency methodology.

3.3 Model integrations

Because this work also needs to be relevant to the shorter-range forecasts made at ECMWF (the present study is a work of opportunity), the model resolution used here is T159 (approximately 1.125° latitude / longitude grid) with 60 levels and with a timestep of 30 minutes. In principle the method should also work at lower resolutions such as at T42.

Analyses at ECMWF are created using 4 dimensional variational data assimilation (4D Var). 4DVar involves the use of the non-linear forecast model and its tangent-linear version to combine "first guess" information from a previous forecast with the new observations. Here a separate set of analyses is made for each model version. Analyses are produced every 6 hours from 00 UTC on 27 December 2004 to 18 UTC on 31 January 2005. Note that the very first analysis (00 UTC on 27 December 2004) is based on a first guess from the operational forecast. Subsequent analyses are based on first guess forecasts using the consistently perturbed model. For some extra computational expense, the first few analysis cycles could have been discarded to reduce any influence of the initial operational first guess. However, any trends associated with the change in the analysis model should be reflected in the estimated confidence intervals and these are acceptably small for the present study. The computational costs involved in making the analyses and forecasts are discussed later.

For each model version, five-day forecasts are started every six hours from 0) UTC on 27 December 2004 to 18 UTC on 26 January 2005, using the corresponding analyses as initial conditions. (Thus the verification times for a D+5 forecast correspond to the whole of January 2005 exactly).

For completeness, climate experiments are also run with the CONTROL and ENTRAIN/5 model. These are explained later. Note that for the initial tendency methodology to be useful, it is almost certainly vital that the initial tendency experiments are conducted at exactly the same spatial and temporal resolution as that used for the climate forecast experiments themselves.

4. NWP skill

Before assessing initial tendencies, we investigate how well these models perform in conventional NWP mode. Table 2 shows D+5 mean spatial anomaly correlation coefficients (ACCs) and root-mean-square errors (RMSEs) for northern hemisphere 500 hPa geopotential heights (Z_{500}) and tropical 850 hPa temperatures (T_{850}) based on the 12 UTC forecasts. Also shown are corresponding values for the operational high-resolution (T511) forecast (termed "OPERATIONS"). It is clear that OPERATIONS produces the best weather forecasts (it is always statistically significantly better than CONTROL at the 5% level, as signified by the "+"'s in the table). CLOUD is not significantly different from CONTROL. ENTRAINx3 is not significantly worse than CONTROL in the northern hemisphere although it is worse in the tropics. ENTRAIN/5 is significantly worse than CONTROL for all scores shown.

Comparisons like the one above are performed routinely at NWP centres to assess possible system developments. However comparison of NWP skill scores is not a very direct or reliable method of assessing

the physics of climate models. For example, CONTROL has exactly the same physics as OPERATIONS and may, therefore, produce an equally good climate even though it is worse at weather prediction. For models with the same resolution (e.g. CONTROL and ENTRAIN/5) there may be more justification in rejecting or down-weighting the model with the worse NWP scores. Nevertheless, the NWP skill test remains a very indirect method of assessing model physics and is computationally more costly than the initial tendency method that only requires the simulation of a few model timesteps. Importantly, it is unclear how the NWP forecast skill for two or more differently perturbed models could be combined to give an estimate of the skill of the model that includes both sets of perturbations. Without this "linearity", the assessment of every differently perturbed model would require a new set of analyses and forecasts to be made and the cost of model assessment would be prohibitive.

Table 2 Day+5 skill scores for each model

Model	Resolution	N. Hem Z_{500}		Tropical T_{850}	
		ACC	RMSE (m)	ACC	RMSE (K)
OPERATIONS	T511	0.90 (+)	497 (+)	0.74 (+)	1.03 (+)
CONTROL	T159	0.87	552	0.69	1.19
CLOUD	T159	0.87	558	0.68	1.19
ENTRAIN/5	T159	0.85 (-)	591 (-)	0.62 (-)	1.32 (-)
ENTRAINx3	T159	0.87	560	0.65	1.48 (-)

Mean day+5 spatial anomaly correlation coefficients (ACCs) and root-mean-square errors (RMSEs) for Northern Hemisphere 500 hPa geopotential height (north of 20oN) and Tropical 850 hPa temperature (between 20°S and 20°N). "+" ("-") indicates the mean score is statistically significantly better (worse) than the corresponding value for the CONTROL model at the 5% level using a paired 2-sided t-test taking autocorrelation into account. The climatology used in the calculation of the ACCs is ERA-40.

5. The use of initial tendencies to assess model physics

The systematic initial tendency is the sum of all the individual tendencies associated with the dynamical and physical processes represented in the model. Hence for simplicity and to avoid confusion, we refer to the "systematic initial tendency" (in a state variable) as the "total initial tendency" or simply the "total tendency". If the total tendency is not zero, there is an imbalance between the individual process tendencies. To illustrate how the initial tendency methodology can be used to understand model error we have calculated vertical profiles of the initial tendency budgets over the Amazon/Brazil region (300°E-320°E, 20°S-0°N). We could have chosen any region over-which the balance of the dominant physical processes is thought to be reasonably uniform. However, if we had accidentally chosen a region where CONTROL happens to represent this balance well, then our results could have been biased towards the CONTROL model. We chose the Amazon/Brazil region because the CONTROL model actually has a problem with the climate in this region at this time of year: the convection "spins-down" quite strongly so that even by D+5 the precipitation rate is on average about 80% of what it should be (see Table 3 later). Hence the choice of this region should not favour the CONTROL model in the present study (and is useful for future model development). Note that our final conclusions will be based on averages over the entire globe.

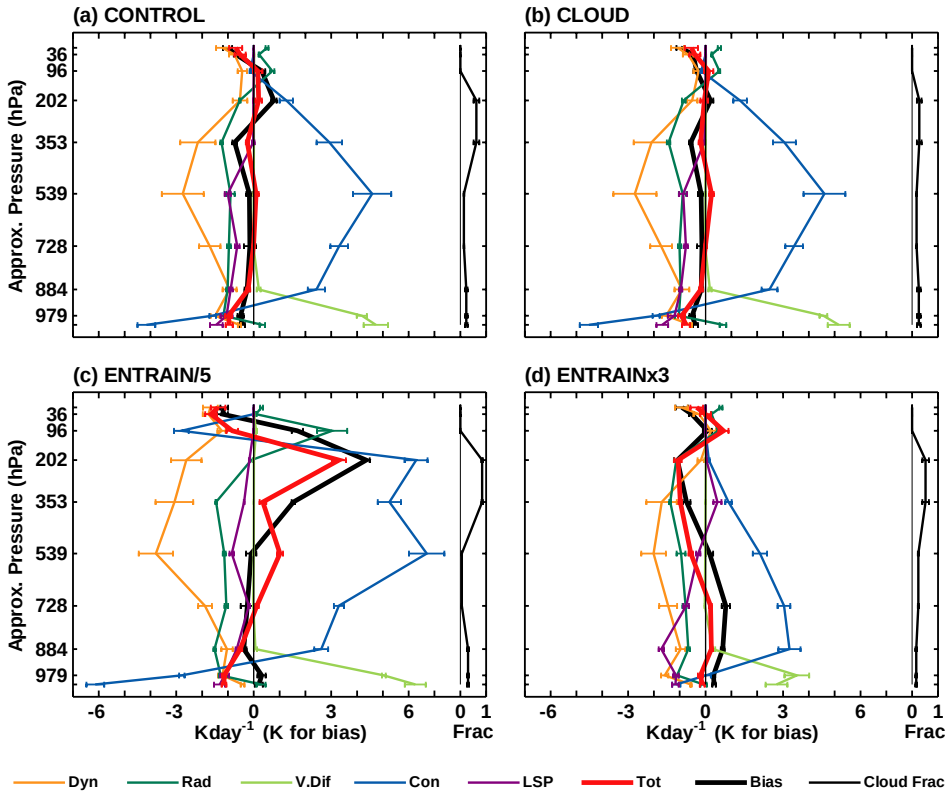

Figure 4 Vertical profiles of initial temperature tendencies for the Amazon/Brazil region based on January 2005 forecasts for (a) the CONTROL model, (b) the CLOUD model with modifications to the large-scale cloud, (c) the ENTRAIN/5 model with reduced convective entrainment and (d) the ENTRAINx3 model with increased convective entrainment. The initial tendencies shown are indicated in the key and correspond to each model's dynamical tendencies, the tendencies from each of the dominant physical processes and the total tendency. Also shown are vertical profiles of the D+5 systematic error (bias) and the cloud fraction (right-hand profile in each panel). The bars indicate 70% confidence intervals. Mean tendencies are calculated on every 5th model level. The vertical coordinate is linear in pressure and represents the approximate pressure at these model levels. See the main text for more details.

Fig. 4a shows the CONTROL total temperature tendency (thick red) and the individual tendencies from the dominant processes (thin variously coloured). The dominant balance is between convective warming due to latent heat release (blue) and dynamic cooling due to ascent (orange). The radiative destabilization of the profile (dark green) is crucial for triggering the convection and is also important in the overall balance. Evaporation in down-drafts leads to a net cooling by the large-scale cloud scheme (purple). The "vertical diffusion" scheme (light green), which includes the surface sensible heat flux, is important below 884 hPa. The total tendency (thick red) is small in comparison to the individual terms; suggesting that the model is in 'reasonable' balance. Nevertheless, there is a net cooling below 884 hPa, a small net cooling at around 353 hPa and a small net warming between about 202 and 96 hPa.

For reference, Fig. 4b shows the corresponding profiles for CLOUD. The modifications that go into CLOUD tend to decrease high cloud cover and increase medium-height cloud cover (compare the cloud fraction profiles in Fig. 4a and 4b) and increase cloud ice (not shown). The main consequence is more radiative cooling above 539 hPa. The difference is most apparent at 202 hPa (compare the dark green curves in Fig. 4a and 4b). At 202 hPa the difference is -0.30 ±0.03 $Kday^{-1}$. (Note that the uncertainty represents the 70% confidence interval using a 'paired' t-test applied to the difference timeseries: it is considerably smaller than

the individual 70% confidence intervals). The change is seen to improve the initial total temperature tendency (thick red) throughout the troposphere (except for at 539 hPa) and, at D+5, temperature biases (thick black) are also improved. Unfortunately there is a slight worsening of the initial total moisture tendencies (not shown). Table 3 shows the precipitation rate and total cloud fraction for each experiment at D+5. The increased thermal destabilization in CLOUD of the upper-tropospheric profile leads to a better mean precipitation rate: up from 5.8 to 6.6 mmday^{-1} (+0.8 ±0.1 mmday^{-1}), with the observed value for January 2005 being 7.1 mmday^{-1} (based on "GPCP" monthly precipitation analysis, Adler et al., 2003). However the total cloud fraction at D+5 is worse: down from 0.78 to 0.60 (-0.17±0.02), with an observed value of 0.77±0.04 (based on "MODIS" monthly cloud cover analysis, see e.g. Platnick et al., 2003). Hence the results may lead one to hypothesize that we need to improve the radiative impact of cloud rather than modify the total amount of cloud.

These CLOUD versus CONTROL results demonstrate the magnitude of a change in model imbalance associated with a "typical" NWP development. One could argue that this magnitude is a measure of our uncertainty in the representation of the physical processes. As such, one would hope that climate model perturbations should result in roughly similarly-sized changes to the total tendencies. However, as we will see below, the perturbations used here result in much larger changes.

Table 3 Precipitation and cloud fraction for each model

Data source	Precipitation (mmday^{-1})	Cloud Fraction
OBSERVATIONS	7.1	0.77±0.04
CONTROL	5.8±0.9	0.78±0.07
CLOUD	6.6±1.2	0.60±0.11
ENTRAIN/5	6.1±0.4	0.75±0.06
ENTRAINx3	8.6±2.3	0.75±0.06

January 2005 observed and forecast (at D+5) precipitation rate and cloud fraction for the Amazon/Brazil region. Observed precipitation rate is based in GPCP data and observed cloud fraction is based on MODIS data. 70% confidence intervals are given for forecast values. The confidence interval for observed cloud is based on the difference between ISCCP (Rossow and Schiffer, 1991) and MODIS cloud fraction for January 2001.

Fig. 4c and d show the corresponding profiles for ENTRAIN/5 and ENTRAINx3, respectively. In the ECMWF model version used here, entrainment of moisture into a convective plume and detrainment of moisture out of a convective plume are composed of an organized part and a turbulent part. Organized entrainment occurs in the lower part of the cloud and is proportional to the large-scale dynamic convergence of moisture. Organized detrainment is related to the vertical variation of the up-draught vertical velocity. The turbulent entrainment rate of environment air and detrainment rate of updraught air are both proportional to the up-draught mass-flux. In ENTRAIN/5 and ENTRAINx3, we perturb the constant of proportionality for the turbulent component. With these changes, there is (initially) deeper and more vigorous convective heating (blue) for ENTRAIN/5 (Fig. 4c) compared to CONTROL (Fig. 4a). This is consistent with reduced loss of moisture and buoyancy from the convective plume with decreased detrainment (assuming that the observations constrain the analysis sufficiently so that the analysis using the ENTRAIN/5 model is as moist as that using CONTROL). On the other hand, ENTRAINx3 initially shows weaker and shallower convective heating (Fig. 4d, blue) than CONTROL (Fig. 4a). This is consistent with increased loss of moisture and buoyancy from the plume with increased detrainment. Importantly, for both ENTRAIN/5 and ENTRAINx3, the changes in convective tendencies are reflected in larger magnitudes of the total initial tendencies (thick red) compared to CONTROL. The temperature of ENTRAIN/5, and to a lesser extent ENTRAINx3, is

clearly more out-of-balance at the analysis state. We argue, therefore, that these two perturbations represent a degradation of the physics of the base model for this particular parameter (i.e. temperature), region and time of year.

Notice that the initial dynamical tendencies (Fig. 4, orange) differ greatly from one model to the next. ENTRAIN/5 (Fig. 4c) shows strong dynamical cooling in the mid-troposphere, consistent with strong large-scale low-level convergence. On the other hand, ENTRAINx3 (Fig. 4d) shows much weaker dynamical cooling. This difference in initial dynamical tendencies indicates that the observations are not able to perfectly constraint even the large-scale analysed flow. Initiation of model ENTRAIN/3 with analyses generated using model ENTRAIN/5 would probably result in an initial "spin-down" of the large-scale flow. The tendency associated with this spin-down in the ENTRAINx3 forecast could not be totally attributed to physics errors in the ENTRAINx3 model. Similarly, the mean analysed specific humidity over the Amazon/Brazil region at about 728hPa, is 7.2 gkg^{-1} when the CONTROL model is used in the data assimilation (not shown) but only 6.0 gkg^{-1} when the ENTRAIN/5 model is used. Initiating ENTRAIN/5 with analyses generated using CONTROL could result in an initial loss of moisture that may not be totally attributable to errors in ENTRAIN/5 physics. Such analysis differences do not invalidate the initial tendency approach (which simply requires that the observations are able to partially constrain the analysis) but rather emphasize the importance, for a fair comparison of models, of generating analyses with a consistent model.

By D+5, balance is (nearly) achieved in ENTRAIN/5 and ENTRAINx3 (i.e. the model has "spun-up" so that the systematic total tendency at D+5 is approximately zero). However an explanation of the balanced budgets (not shown) is much more complicated than the explanation of those in Fig. 4 because other processes (including the large-scale dynamics) have had time to interact with the perturbed convection scheme. Hence, although D+5 biases (thick black) also clearly indicate model physics errors (there is a good agreement between the profiles of total initial tendency and D+5 bias), it would be difficult to identify the cause of this error based on the D+5 biases alone. Importantly in terms of this study, the D+5 precipitation rates and cloud cover fractions (Table 3) are very different from what one would naively expect from the initial tendencies and, in fact, not noticeably worse than that of CONTROL. This compensation suggests that considerable interaction between the various model processes has already occurred by D+5.

6. Scoring climate models with initial tendencies

Having demonstrated the power of the initial tendency method to assess model physics within a particular region, we now consider the global picture. Fig. 5 shows fields of vertically-integrated *absolute* initial tendencies of temperature (top) and specific humidity (bottom) for CONTROL (left) and ENTRAIN/5 (right). We use the absolute values of the tendencies so that if oppositely signed tendencies exist at different levels, these will not cancel each other. It is clear that initial temperature tendencies for ENTRAIN/5 are much worse than for CONTROL throughout the tropics and subtropics and moisture tendencies are much worse in the Inter-Tropical Convergence Zone and monsoon rainfall regions (which are south of the equator in January). Globally-integrated absolute initial tendencies of T, q, u and v for each model are given in Table 4. While ENTRAIN/5 is poor for T, q, u and v, ENTRAINx3 is only poor in terms of T when averaged over the globe. CLOUD shows little difference from CONTROL. Hence we can conclude that the physics of CONTROL are much better than those of ENTRAIN/5, better than those of ENTRAINx3 and approximately equivalent to those of CLOUD.

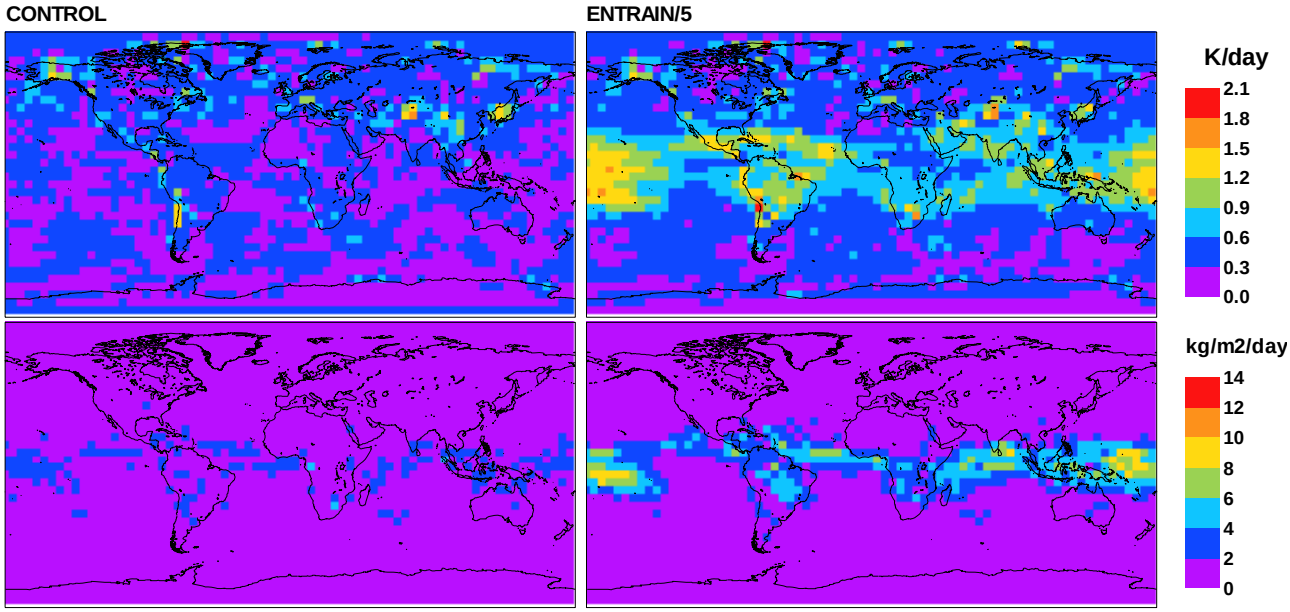

Figure 5 Vertically-integrated absolute total tendencies of (top) temperature and (bottom) specific humidity for (left) the CONTROL model and (right) the ENTRAIN/5 model with a reduced turbulent entrainment coefficient. Vertical integrals are mass-weighted integrals of the absolute values of the total tendency based on a subset of 11 model levels ranging from the surface to about 4 hPa.

Table 4 Globally and vertically integrated absolute initial tendencies

Model	dT/dt ($Kday^{-1}$)	dq/dt ($kgm^{-2}day^{-1}$)	du/dt ($ms^{-1}day^{-1}$)	dv/dt ($kgm^{-2}day^{-1}$)
CONTROL	0.378	1.03	1.26	1.23
CLOUD	0.376	1.03	1.25	1.22
ENTRAIN/5	0.608	1.73	1.52	1.48
ENTRAINx3	0.454	1.02	1.25	1.24

Vertical integrals are mass-weighted integrals of the absolute values of the total tendency based on a subset of 11 model levels ranging from the surface to about 4 hPa. Global integrals are area-weighted.

A single (scalar) error score for each model would be very useful to assess each model's representation of the physics. Various methods have been explored. A particularly simple score can be based on the mean absolute tendency. To place equal weight on each of the four parameters (T, q, u and v), the integrated absolute tendencies in Table 4 need to be first normalised (we do this by dividing by the mean value over all models because simple division by the values for CONTROL could lead to biases). After normalisation, the mean over all parameters is made. The scores for each model based on the absolute tendencies are CONTROL: 0.90, CLOUD: 0.89, ENTRAIN/5: 1.27, ENTRAINx3: 0.94. The lower this score is, the smaller are the initial tendencies, and the better is the model. The score tends to zero in the limit of a large ensemble of tests and a perfect forecasting system.

7. Linearity of the initial tendency methodology

Murphy et al. (2004) assumed that the anomalous climates from perturbed models can be combined linearly. This assumption was essential to make the computational overhead of the problem of producing a p.d.f. of climate change manageable. About 23 of their parameters were associated with fast physics. As indicated in the representative Table 1, these parameters either took low, medium and high values or were associated with on/off switches (or a combination of both). The total number of possible combinations of these

parameter perturbations is about 15 billion ($=3^5 2^2 .4^2 3^2 .3^1 2^3 .3^3 2^1 .3^4$). If we could assume that anomalous climates combine linearity, then just 24 models would be needed to span this combination space (23 single perturbations plus the control model). Whilst linearity of climates is an appealing concept, it is unclear whether it is a scientifically justified assumption to make; indeed, our D+5 climate and budget analysis casts doubt on this assumption.

On the other hand, as discussed below, we believe that the initial tendency method is "linear enough" to allow major computational savings to be made. Further perturbations have been applied to the control model to reach this conclusion. One such perturbation involves a doubling of the ice particle radius (throughout the size distribution). We call this model ICEx2. Another model incorporates both the ENTRAINx3 and ICEx2 perturbations. One could imagine that the doubling of the effective radius of ice particles in ICEx2 will primarily affect the initial radiation budget whereas a perturbation to the entrainment rate in ENTRAINx3 will primarily affect the convection. Hence initially these two perturbations may not interact too strongly and thus their combined effect in a single model may be expected to be approximately the same as the sum of their effects when applied individually. If, for the initial tendency anomalies we write E = ENTRAIN/3 - CONTROL, I = ICEx2 - CONTROL and EI = (ENTRAINx3 and ICEx2 combined) - CONTROL, then the degree of linearity can be assessed by calculating the magnitude of the nonlinear residual: EI-(E+I). Vertical profiles of the anomalous tendencies have been calculated for the same Amazon/Brazil region but the magnitudes associated with E are so much greater than those for I that linearity is trivially true: EI-(E+I)≅E-(E+0)=0. To find a region where both E and I are comparable in magnitude, the latitude circle at 60oS was chosen. During January, there is plenty of sunlight at 60°S which should favour I but there is less deep convection than in the South American monsoon region and this should reduce the impact of E.

Figure 6 shows the vertical profiles of anomalous total initial temperature tendencies averaged around the 60oS latitude circle for E, I, EI and EI-(E+I). Within the troposphere the non-linear component, EI-(E+I), is not significantly different from zero and generally has a smaller magnitude than the magnitudes of E or I alone. This suggests a reasonable degree of linearity.

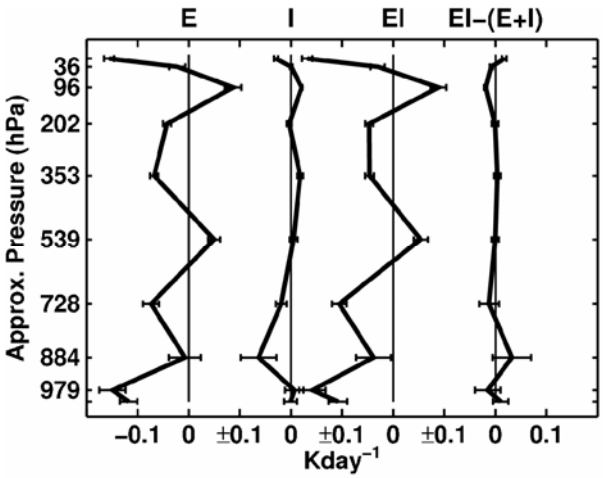

Figure 6 Vertical profiles of total initial tendency anomalies at 60oS for the high convective entrainment model E, the model with increased ice particle size I, the model with both these perturbations EI and the non-linear term EI-(E+I). The bars indicate 70% confidence intervals. See the main text for more details.

It should be noted however, that contrary to the speculation above, it is not just the initial convective tendency in E or the initial radiative tendency in I that are affected by the perturbations applied. For example, the upper tropospheric dynamical tendency anomalies in E and I are statistically significantly different from

zero. The tendency differences from processes other than the process which contains the perturbed parameter could arise through the analysis cycling or as a direct response in the 6-hour forecast to the applied perturbation. Whatever the reason, the degree of linearity displayed above suggests that these processes respond approximately linearly to the applied perturbation, and these responses are not strong enough to interact greatly during the first few timesteps. In the lower troposphere, large confidence intervals for some of the individual tendency anomalies (not shown) suggest that more than 31 days are required to adequately estimate these individual terms, at least in the extratropics.

For a more general and global perspective on linearity, further initial tendency experiments have been made with a perturbation to the critical relative humidity for cloud formation. This perturbation is applied on its own and in combination with ICEx2 and with ENTRAIN/3. The corresponding anomalous tendencies and denoted R, IR and RE, respectively and the non-linear terms are calculated as IR-(I+R) and RE-(R+E). The globally and vertically integrated absolute values of the anomalous and non-linear tendencies are shown in Table 5. In general, the magnitude of the non-linear term for a given parameter is approximately the same for all sets of experiments and is generally equivalent to the magnitude of the smallest individually perturbed model. Given that there could be some inflation of the non-linear terms due to larger sampling errors than for the anomalous tendencies, it is clear that a useful degree of linearity is present throughout the globe in the initial tendency methodology. For example, although the linearity may not be perfect, we should be able to estimate the absolute tendency score for the combined perturbation models from those of the singly perturbed models. This is what we will mean by 'linearity' from now on. Note that the linearity found for these experiments is not a trivial consequence of the different perturbations affecting disjoint geographical regions since E, I and R all have their strongest anomalies in the same locations (the heavily convective regions). Nevertheless, it is possible that other perturbations not considered here may affect disjoint regions and the trivial linearity that would ensue would still be highly useful.

Here, we have only assessed linearity for fixed perturbation values. Given the non-linearities inherent in some physical parametrizations, it is perhaps doubtful that the response to a single perturbation will vary linearly with the magnitude of that perturbation. For each of the (n) perturbable parameters, one could imagine determining the initial tendency response to a range of perturbation values (for example by making 3 low/medium/high value simulations and fitting a curve through the three responses or by making 2 on/off simulations for switch-like parameters). The responses of these ~3n single-parameter perturbed models could therefore be used to predict (approximately) the initial tendency response for any linear combination of parameter perturbations. The vertical and global integration of the absolute values of the linear combination of initial tendencies could then be used to assess the physics of the corresponding model. A possible extension of this methodology may be required for the situation where two (or more) perturbations are applied within the same physical process if these perturbations can interact with each other in a single call to the routine. In the situation of m interacting perturbations in a given process, $\sim 3^m$ (rather than ~3m) perturbed models would require assessment, corresponding to all combinations of low, medium and high (or on/off) for each parameter. Even with this extension, the 'inter-process linearity' highlighted above would represent a considerable reduction in the effort needed to assess the model physics with the initial tendency approach. For example, the 23 parameters associated with "fast physics" used by Murphy et al (2004) are distributed over 5 physical processes and would require initial tendencies to be simulated for 1275 ($=3^5 2^2+4^2 3^2+3^1 2^3+3^3 2^1+3^4$) perturbed models. This is very much less than the ~15 billion simulations required in the fully non-linear situation and, considering the importance of trying to reduce uncertainty in estimates of climate change, may not be out of the question computationally.

Table 5 Globally and vertically integrated absolute initial tendency anomalies

Anomaly	dT/dt (Kday^{-1})	dq/dt (kgm-2day^{-1})	du/dt (ms^{-1}day^{-1})	dv/dt (kgm^{-2}day^{-1})
E	0.19	0.36	0.25	0.20
I	0.03	0.09	0.05	0.05
R	0.02	0.09	0.04	0.04
EI	0.20	0.36	0.26	0.21
IR	0.03	0.09	0.05	0.05
RE	0.19	0.36	0.25	0.20
EI-(E+I)	0.03	0.13	0.06	0.06
IR-(I+R)	0.03	0.12	0.05	0.05
RE-(R+E)	0.03	0.13	0.06	0.06

The mass-weighted vertical integration is between approximately 728hPa and 36hPa. See the main text for further details.

8. Computational costs

At the resolutions used in this study (T159 horizontal resolution, 60-level, 30-minute timestep), the cost of the data assimilation represents 98% of the cost of the initial tendency method. The total cost of a 31-day, 4-times-a-day, 6-hour initial tendency analysis (including data assimilation and forecast) is equivalent to 4.7 years of coupled model simulations at the same resolution. If only 100 years of simulations are required to spin-up a coupled model and assess its present-day climate, then the initial tendency method will cost just 4.7% the cost of the present-day climate test. However, it is not unusual to discard the first 300 years of a coupled model simulation prior to assessing its climate. These relative costs should be approximately independent of model resolution as long as the number of observations ingested by the data assimilation system is proportional to the resolution. However, since the initial tendency method will be more discriminating at higher resolution, where more information from the available observations can be assimilated, the proposed technique should perhaps be considered in conjunction with the development of the next generation of higher-resolution climate models.

9. Using initial tendencies within perturbed model climate ensembles

The aim of the present study is to assess the representation of fast physics within climate models using a reasonably small number of data assimilation / forecast cycles and integrating tendencies over the diurnal cycle. The integrated absolute tendency score was introduced in section 6 as a means of assessing model physics. One may wonder whether this score could be used to weight models within an ensemble of model versions. For example, weights could be based on the reciprocal of the absolute tendency score (appropriately scaled so that the weights sum to 1). The weights for the models assessed in section 6 would be CONTROL: 0.27, CLOUD: 0.28, ENTRAIN/5: 0.19, ENTRAINx3: 0.26. If there were a perfect model within the ensemble then, as the number of assessment forecasts tends to infinity, its absolute tendency score should tend to (approximately) zero and its weight would tend to 1 with all other models' weights tending to zero. This seems appealing although it is clearly somewhat arbitrary to attach a weight of 0.5 to a model that has twice the initial tendency. Why not base the weight on the reciprocal of the mean *squared* tendencies for example?

It is intuitively appealing to try to estimate, for a given finite number of assessment forecasts, the probability that each model could be "perfect". Although it is not trivial to construct such a probability, one possible method is outlined here. We have produced such a probability score by estimating for each grid-point, model level and parameter, the probability that a 31-day sampling of a *perfect* model could lead to a mean tendency

whose magnitude is as large as that of the actual mean tendency calculated from the assessment forecasts. Explicitly, if $\{\dot{T}\}$ is the timeseries of daily-mean tendencies for a given model, grid-point, level and parameter and $\bar{\dot{T}}$ is the mean of this timeseries, $S_{\dot{T}}$ its standard deviation and ρ its lag-1 autocorrelation then, assuming normality, we can solve for the largest probability P such that:

$$t_{N'-1,1-P/2} \geq \left| \frac{\bar{\dot{T}}}{\frac{1}{\sqrt{N'}} S_{\dot{T}}} \right| \quad ,$$

where $t_{N'-1,1-P/2}$ represents the percentiles of the Student's t-distribution with $N'-1$ degrees of freedom. N' is at most the number of days in the sample (here 31) but can be smaller if $\rho > 0$ (see von Storch and Zwiers; 2001, for further details). As with the absolute tendency score, we would like to produce some average probability that the sample means are consistent with a perfect model. To do this, we make a mass-weighted vertical average of P and then also average over the four parameters (T, q, u, v). Experimentation shows that an area-weighted global average tends to 'dilute' the impact of a poor representation of the tropical physics in ENTRAIN/5. It seems natural to assume that different physical errors may be dominant in the tropics (20°S-20°N) and the extratropics (beyond 20° latitude in both hemispheres) and so here we integrate over the tropics and extratropics separately to produce two area-mean probabilities: P_{TROP} and P_{EXTR}. Assuming independence of errors, the product $P_{TROP} \times P_{EXTR}$ can be considered to be the probability that the sample means in both regions are consistent with a perfect model. The probability scores derived in this way are CONTROL: 0.202, CLOUD: 0.198, ENTRAIN/5: 0.122, ENTRAINx3: 0.197. The probabilities for all models are low; indicating that none of the results is particularly consistent with a perfect model hypothesis. The probability for ENTRAIN/5 is considerably lower than the rest, which seems reasonable based on the results of sections 5 and 6. By scaling these numbers to sum to one, we obtain "probability weightings" for the models that could be used within a multi-model ensemble: CONTROL: 0.28, CLOUD: 0.27, ENTRAIN/5: 0.18, ENTRAINx3: 0.27. Although other choices could have been made in our definition of the probability weighting, it is interesting to note that these weightings are very similar to those of the reciprocal of the absolute tendency score (above). Because the probability assessment involves the standard deviation of the estimated mean, $S_{\dot{T}}/\sqrt{N'}$, the probability weightings will be dependent on the number of assessment forecasts made. The more computational time spent assessing model physics, the more discriminating the weights will be. As with the absolute tendency score, the probability weighting that would be attached to a perfect model (if one existed in the ensemble) should tend to 1 as the number of assessment forecasts increases.

The practical application of the initial tendencies methodology to climate sensitivity may, therefore include:

1. Initial expert opinion on reasonable maximum magnitudes for each parameter perturbation.
2. Estimation of the probability weightings for each perturbation using the initial tendency approach.
3. Possible reduction in some perturbations if they are very unlikely (i.e. have very low probabilities). Possible increase in some perturbations if their probabilities remain too high (so that parameter space is adequately sampled). Return to (2) if necessary.
4. Calculation, using approximate linearity, the probability weightings for multi-perturbed models.

As discussed in section 2, the initial tendency approach focuses only on 'fast physics' errors and additional tests are required to assess other aspects of the physics. Subsequent weights such as from tests dedicated to

assessing the salient 'slow physics', and from present-day climate tests, low frequency variability tests, palæoclimate tests (Annan et al., 2005a) and from tests of the physics most directly related to the greenhouse effect would need to be combined with the weights from the initial tendency test. Just how this combination is done is an important question but is not explored fully here. One pragmatic possibility might be to make the ultimate weight a suitably normalised minimum of all these weights but other combination methods or Bayesian approaches could be explored. After the combined weights have be decided on, the climate forecasts can be made and a p.d.f. of climate change constructed.

10. Climate experiments

It is interesting to check whether ENTRAIN/5 would pass or fail a standard present-day climate test. To do this 39 17-month simulations were made with CONTROL and ENTRAIN/5 starting from 1 October for the years 1962-2000. The initial data coming from the ECMWF reanalysis project (ERA-40). The models were forced with prescribed (observed) sea-surface temperature (SST). Comparison of months 3-5 with months 15-17 (both corresponding to the December - February season) indicated that the models were well "spun-up" by month 3. For each simulation, the last 12 months (months 6-17) were then used to assess the model climate. Our present-day climate assessment is similar to that used by Stainforth et al. (2005). The assessment was based on the simulation of ERA-40 MSLP and T850 for the years 1962-2000 and GPCP precipitation for the years 1979-2000. T_{850} was used instead of 2m temperature (T_{2m}) because T_{2m} is likely to be too closely related to the prescribed SST. The present-day climate error for ENTRAIN/5 was 1.39 (blue diamond in Fig. 3), relative to the value of 1.00 for CONTROL (red diamond in Fig. 3). It is evident from Fig. 3 that the present-day climate error of ENTRAIN/5 is less than the errors of some of the models accepted by Stainforth et al. (2005) and in this respect we can conclude that ENTRAIN/5 passes the present-day climate test. The passing of a present-day climate test by a model that represents the physics of the atmosphere so poorly, highlights the importance of having other means (such as the initial tendency method) of down-weighting or removing bad models from climate forecasts.

The climate sensitivity of CONTROL and ENTRAIN/5 to a doubling of CO_2 is also indicated by the diamonds in Fig. 3. For both models, the climate sensitivity is very small and of little interest due to the prescribed nature of the SSTs. Nevertheless it is interesting to note that the top-of-atmosphere fluxes of heat associated with the enhanced greenhouse effect are similar to those of other (coupled) models. The *pattern* of surface warming, with largest temperature increases over the northern hemisphere continents, is also in agreement with coupled model forecasts. The effectively-infinite heat capacity of the oceans means that considerable heat is removed from the system at the surface, particularly in the northern hemisphere stormtrack regions. The representation of these surface fluxes is likely to be critical to the climate sensitivity of transient CO_2 experiments and could also be assessed with the initial tendency methodology.

11. Discussion and conclusions

Forecasts of anthropogenically-forced climate change remain highly uncertain. Recent attempts to systematically account for the uncertainty associated with model error (using perturbed model ensembles) have, if anything, increased this uncertainty. New approaches are required to constrain our p.d.f.s of climate change in order to guide strategic decision making on mitigation and adaptation.

This study has developed an approach that can be used to quantify better the uncertainty in climate change forecasts due to model error. The methodology uses very short-range numerical weather prediction (NWP) to calculate the imbalance of the climate model about a realistic atmospheric state (the analysis). This

imbalance is a manifestation of errors in the model's "fast physics" and can therefore be used to produce weightings that reflect how well each model represents this physics. Note that these fast physics errors include not only errors in represented physics but also errors due to unrepresented processes. The methodology would actually result in reduced (not just more accurate) climate change uncertainty if the models with the poorest representation of the physics have the highest (or lowest) climate sensitivities. Indeed, the model *perturbation* that led to the highest climate sensitivity in Stainforth et al. (2005) and which was the basis for "plausible" global warming quotes in the press of 12°C (see e.g. Pearce 2006) is rejected as unrealistic by the present methodology (although it must be stressed that we use a different base model and cannot, therefore, use our result to constrain their range of climate sensitivity).

The methodology involves the calculation of tendencies over the first few timesteps of a model forecast. Computationally, the methodology is more efficient than assessing a model's ability to simulate present-day climate. Importantly, it has been demonstrated that these initial tendencies combine in a near linear fashion (due to linearity and/or spatial orthogonality in the effects of different perturbations). Hence it is possible that the fast-physics of a model with any combination of parameter perturbations can be approximately assessed from the initial tendencies associated with each individual perturbation alone. This approximate linearity could further reduce the computational expense of assessing model fast physics by several orders of magnitude and thus help greatly in the quantification of climate change uncertainty.

The initial tendency methodology as presented here has strong similarities with the simple quantification of "analysis increments". The main advantage of the initial tendency approach is the availability (for a 2% increase in cost) of the tendencies from each individual physical process: a feature that is essential for demonstrating physically the utility of the approach. The initial tendency approach also has the advantage (important here) that the first timestep of the forecast can be disregarded whereas it is used implicitly in the calculation of the analysis increment. Clearly the precise approach that is most applicable may be dependent on the base climate model under investigation.

By fitting a set of simple climate models (each with differing physics packages) to the observational record, Andronova and Schlesinger (2001) also found that the climate sensitivity was highly dependent on the physics represented in the model. They found very high climate sensitivities (over 10K) were possible but only if the total effect of anthropogenic sulphate aerosol emissions has strongly offset greenhouse warming up until now. The present initial tendency methodology has also been used successfully to discriminate between different model aerosol climatologies (Rodwell and Jung 2007) although the focus was on the direct effect of desert aerosol. If the initial tendency methodology could be used to assess the direct and indirect effects of anthropogenic sulphate aerosol emissions (in regions where these emissions are high) then this could represent an important step forward in climate change forecasting.

Other applications of the initial tendency methodology may be in assessing ocean-atmosphere interactions in atmospheric models and the representation of slower oceanic processes in coupled models. Both these applications would be relevant to seasonal forecasting as well as to climate forecasting.

Above all, this study highlights the importance of developing a seamless approach to weather and climate forecasting, where we glean as much information as we can from our daily weather observations in order to better predict the climate for centuries ahead. (A seamless approach could be a two-way process with the identification and correction of 'slow-physics' errors in long-range forecasts also leading to improvements at shorter timescales). This seamless approach is consistent with the strategic plan of the World Climate Research Programme (WCRP COPES 2005). In a truly seamless system in which a climate model can be run

in data assimilation mode, uncertainty in climate sensitivity, one of the long-standing problems in climate change research, may be reduced significantly. By contrast, if seamless systems are not developed, it will not be possible to use the results of the proposed method in any quantitative way.

Acknowledgments

The authors would like to thank David Stainforth for his important help in the initial stages of this study and for the data presented in Fig. 3. Thanks also to Thomas Jung, Mike Fisher, Lars Isaksen and Adrian Tompkins for useful discussions, and the reviewers for their valuable suggestions.

References

Adler, R. F., and Co-authors, 2003: The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-present). *J. Hydrometeor*., **4**, 1147-1167.

Andronova, N. G. and M. E. Schlesinger, 2001: Objective estimation of the probability density function for climate sensitivity. *J. Geophys. Res.* **106**, 22605-22612.

Annan J. D., J. C. Hargreaves, R. Ohgaito, A. Abe-Ouchi, S. Emori, 2005a: Efficiently constraining climate sensitivity with paleoclimate simulations. *SOLA*. **1**, 181-184.

Annan, J. D., D. J. Lunt, J. C. Hargreaves and P. J. Valdes, 2005b: Parameter estimation in an atmospheric GCM using the Ensemble Kalman Filter. *Nonlinear Processes in Geophysics*, **12** (3), 363-371.

Buizza R., M. Miller and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc*., **125**, 2887-2908.

Cubasch, U., G. A. Meehl, G. J. Boer, R. J. Stouffer, M. Dix, A. Noda, C. A. Senior, S. Raper and K. S. Yap, 2001: Projections of future climate change. In: *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* [Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. van der Linden, X. Dai, K. Maskell, C. I. Johnson (eds.)]. Cambridge University Press, 525-582.

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson (Editors), 2001: *Climate Change 2001: The Scientific Basis*. Cambridge University Press. 881pp

Kattenberg, A., F. Giorgi, H. Grassl, G. A. Meehl, J. F. B. Mitchell, R. J. Stouffer, T. Tokioka, A. J. Weaver, and T. M. L. Wigley, 1996: Climate models - projections of future climate. In: *Climate Change 1995: The Science of Climate Change* [Houghton, J. T., L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 285-357.

Klinker, E., and P. D. Sardeshmukh, 1992: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci*., **49**, 608-627.

Mitchell, J. F. B., S. Manabe, T. Tokioka, and V. Meleshko, 1990: Equilibrium climate change. In *Climate Change: The IPCC Scientific Assessment*, Cambridge: Cambridge University Press, 131-172.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768-772.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: a proposal for nonlocal stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc*., **127**, 279-304.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Ann. Rev. Earth Planet Sci.*, **33**, 163-193.

Pearce, F., 2006: *The last generation; how nature will take her revenge for climate change*. Eden Project Books, Random House, London. 324pp.

Phillips T. J., G. L. Potter, D. L. Williamson, R. T. Cederwall, J. S. Boyle, M. Fiorino, J. J. Hnilo, J. G. Olson, Shaocheng Xie and J. John Yio, 2004: Evaluating Parameterizations in General Circulation Models: Climate Simulation Meets Weather Prediction. *Bull. Amer. Meteorol. Soc*., **85**, 1903-1915.

Platnick, S., M. D. King, S. A. Ackerman, W. P. Menzel, B. A. Baum, J. C. Riédi, and R. A. Frey, 2003: The MODIS Cloud Products: Algorithms and Examples from Terra. *IEEE Transactions on Geoscience and Remote Sensing*, **41**, 459-473.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics, *Quart. J. Roy. Meteor. Soc*., **126**, 1143-1170.

Rodwell, M. J. and T. Jung, 2007: The local and global impact of Saharan aerosol. *Q. J. R. Meteorol. Soc*., In preparation.

Rodwell, M. J. and T.N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quart. J. Roy. Meteor. Soc*., **133**, 129-146

Rossow, W. B. and R. A. Schiffer, 1991: ISCCP Cloud Data Products. *Bull. Amer. Meteorol. Soc.*, **72**, 2-20.

Stainforth, D. A., T. Aina, C. Christensen, M. Collins, N. Faull, D. J. Frame, J. A. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. A. Spicer, A. J. Thorpe, M. R. Allen, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403-406.

Tett, S. F. B., P. A. Stott, M. R. Allen, W. J. Ingram, and J. F. B. Mitchell, 1999: Causes of twentieth-century temperature change near the Earth's surface. *Nature*, **399**, 569-572.

Tompkins, A. M., K. Gierens and G. Rädel, 2006: Ice supersaturation in the ECMWF integrated forecast system, *Q. J. R. Meteorol. Soc*., In preparation.

Uppala, S. M., P. W. Kållberg, A. J. Simmons, U. Andrae, V. da Costa Bechtold, M. Fiorino, J. K. Gibson, J. Haseler, A. Hernandez, G. A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, S. Caires, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B. J. Hoskins, L. Isaksen, P. A. E. M. Janssen, A. P. McNally, J.-F. Mahfouf, R. Jenne, J.- J. Morcrette, N. A Rayner, R. W. Saunders, P. Simon, A. Sterl, K. E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo and J. Woollen, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961-3012.

Von Storch, H. and F. W. Zwiers, 2001: *Statistical analysis in climate research*. Cambridge University Press. 484pp.

Wigley, T. M. L. and S. C. B. Raper, 2002: Reasons for Larger Warming Projections in the IPCC Third Assessment Report., *J. Climate*, **15**, 2945-2952.

WCRP COPES, 2005: The World Climate Research Programme Strategic Framework 2005-2015: Coordinated Observation and Prediction of the Earth System (COPES) *WCRP 123*, WMO/TD-No. 1291.