Analysis and Reduction of Systematic Errors through a Seamless Approach to Modelling Weather and Climate.

**Gill Martin**, **Sean Milton**, **Cath Senior**, **Malcolm Brooks**, **Sarah Ineson**

*Met Office, Exeter, UK*

**Thomas Reichler and Junsu Kim**

*University of Utah, USA*

---

*Corresponding author address: Gill Martin, Met Office Hadley Centre, FitzRoy Rd, Exeter, Devon, EX1 3PB, UK.*

E-mail: gill.martin@metoffice.gov.uk
ABSTRACT

The reduction of systematic errors is a continuing challenge for model development. Feedbacks and compensating errors in climate models often make finding the source of a systematic error difficult. In this paper, we show how model development can benefit from the use of the same model across a range of temporal and spatial scales. Two particular systematic errors are examined: tropical circulation and precipitation distribution, and summer land surface temperature and moisture biases over northern hemisphere continental regions. Each of these errors affects the model performance on timescales ranging from a few days to several decades. In both cases, the characteristics of the long-timescale errors are found to develop during the first few days of simulation, before any large-scale feedbacks have taken place. The ability to compare the model diagnostics from the first few days of a forecast, initialised from a realistic atmospheric state, directly with observations has allowed physical deficiencies in the physical parametrisations to be identified which, when corrected, lead to improvements across the full range of timescales. This study highlights the benefits of a seamless prediction system across a wide range of timescales.
1. Introduction

The growth of systematic errors in general circulation models (GCMs) remains one of the central problems in producing accurate predictions of climate change for the next 50-100 years (Randall et al. (2007)). Although great advances in global climate modelling have been made in recent decades (Solomon et al. (2007)), there are still large uncertainties in many processes such as clouds, convection, coupling to the oceans and the land surface (e.g. Cubasch et al. (2001); Koster and team (2004)). There is obviously great interest in predictions of extreme events such as tropical cyclones and how they will change due to anthropogenic forcing. However, tropical prediction remains a particular challenge. Major modes of tropical variability such as the Madden-Julian Oscillation (MJO; Zhang (2005)) and El Niño-Southern Oscillation (ENSO; Allan et al. (1996)) are still relatively poorly understood and often poorly modelled (e.g. Slingo et al. (1996); Sperber and Palmer (1996); van Oldenborgh et al. (2005); Guilyardi et al. (2009)). On shorter timescales the diurnal cycle of convection over tropical land is another key weakness in models (Betts and Jakob (2002); Yang and Slingo (2001)).

One issue impeding progress is that attributing the growth of systematic errors to the modelling of particular physical processes is notoriously difficult in climate GCMs. This is due to both a lack of observational data and to the non-linear interactions amongst various physical processes and the errors in modelling them. A possible way forward is to use short-range (1-5 day) forecasts from a numerical weather prediction (NWP) framework as a means of evaluating parametrisations in climate models (Phillips et al. (2004)). This has a number of advantages. Firstly, the NWP forecasts are run from initial states generated with state of the
art variational data assimilation (e.g. Lorenc et al. (2000); Rawlins et al. (2007)). This means that the errors in the large-scale synoptic flow are minimised and there are no large biases in the circulation due to remote forcing effects (e.g. tropical-extratropical interactions). Such biases in circulation in a climate model make it difficult to determine if the parametrised physical processes are performing poorly because of errors in their formulation or errors in the inputs to the parametrisations themselves. This approach was taken to the extreme limit of examining 1 timestep forecasts by Klinker and Sardesmukh (1992) to identify errors in the momentum balance of the European Centre for Medium-Range Weather Forecasts (ECMWF) forecasts, and by Rodwell and Palmer (2007) to quantify the uncertainty in climate change forecasts due to model error. Secondly, detailed up-to-date observational datasets (e.g. ARM (Ackerman and Stokes (2003)), TOGA-COARE (Webster and Lucas (1992))) can be used to evaluate the individual physical processes at the temporal and spatial scales of individual weather systems.

Conversely, seasonal and climate coupled modelling provide a very strong constraint on the veracity of the global model formulation. The physics-dynamics developed must perform in the coupled ocean-atmosphere-cryosphere environment without systematic drifts (e.g. in Sea Surface Temperatures, SSTs). The longer timescale integrations also allow an evaluation of the model performance in modes of low frequency variability such as ENSO, the Quasi-Biennial Oscillation (QBO; Scaife et al. (2000)), and the North Atlantic Oscillation (NAO; Hurrell and van Loon (1997)). As Hurrell et al. (2009) point out, using similar models for predictions on different timescales can result in improved skill in both weather and climate forecasts through stronger collaboration and shared knowledge among those in the NWP and climate communities working on parametrisations schemes. The Met Office Unified
Modelling system (Cullen (1993)) has benefitted from the synergy between NWP and climate modelling in many past model developments, including a new boundary layer turbulent mixing scheme (Lock et al. (2000); Martin et al. (2000)), orographic parametrisations (Milton and Wilson (1996); Gregory et al. (1998)), and most recently in development of the new semi-implicit, semi-Lagrangian dynamical core (Davies et al. (2005); Martin et al. (2006); Ringer et al. (2006)).

The current operational global NWP and climate configurations of the Met Office Unified Model (MetUM) have very similar dynamical and physical formulations (see section 2). Comparison of zonal mean temperature and zonal wind cross sections (Figure 1) shows that the models have similar systematic errors on the largest scales, despite differences in both horizontal resolution and prediction timescale. This similarity provides the opportunity to tackle such errors through a joint model development approach. The seasonal and decadal modelling frameworks at the Met Office have, until recently, differed from the configuration used for weather and climate timescales. However, new frameworks for seasonal and decadal modelling are currently under development and these will enable the development of systematic errors on the monthly to decadal timescale to be studied. Ultimately it is hoped that the whole suite of MetUM models will benefit from improvements achieved in this way.

In this paper we focus on two particular areas of concern in the MetUM on climate and NWP timescales: (i) tropical performance, in particular excessive precipitation and evaporation over tropical oceans and related circulation errors, and (ii) summer land-surface temperature and moisture biases over northern continents. The paper is arranged as follows: Section 2 outlines the model formulations used in this study. Section 3 discusses the analysis and reduction of the tropical biases and section 4 concentrates on the northern hemisphere
summer land surface temperature and moisture biases. The impact of this work on the overall performance of the models is described in section 5, and conclusions are drawn in section 6.

2. Model Configuration and Experimental Details

The MetUM forms the basis for weather and climate prediction across a wide range of spatial and temporal scales. In this study we focus attention on three prediction timescales.

The first timescale is the short-range (up to 6 days). The operational deterministic global NWP forecast is run twice a day, from 00 and 12 UTC analyses. The analyses are produced using four-dimensional variational data assimilation (Rawlins et al. (2007)), with SSTs and sea-ice initialised once a day during the assimilation cycle then held fixed through the course of the 6 day forecasts. In this study a number of recent and past global NWP model cycles (Table 1) are discussed in the context of improvements to tropical performance and northern hemisphere land surface temperature and moisture biases. Further details of the operational global NWP model cycles are given in Allan et al. (2005, 2007). In addition, 5-day predictions were also run with the climate model configuration initialised from weekly analyses through four recent boreal winter and summer seasons (2001 to 2004). This uses the climate model within the NWP framework following a similar strategy to Phillips et al. (2004).

The second timescale is the medium-range (up to 15 days). Since 2006, the Met Office Global and Regional Ensemble Prediction System (MOGREPS; Bowler et al. (2007a,b)) has been running with a 24-member global ensemble to 6 days ahead on a daily basis. In addition,
ensembles of 15-day forecast runs (MOGREPS-15) are run daily at ECMWF as part of the THORPEX\textsuperscript{1} Interactive Grand Global Ensemble (TIGGE) program (Titley et al. (2008)). Here we utilise results from a series of 15-day deterministic forecasts run as a testbed suite for evaluating new physics parametrisation developments prior to implementation in the MOGREPS-15 production suite (Savage et al. (2006)). The 15-day deterministic forecasts are initialised from Met Office 12 UTC global operational analyses and SSTs are held fixed throughout the 15 days of the forecasts.

The final timescale is the climate predictions. The climate configuration of the MetUM, HadGEM1, is described by Martin et al. (2006) and Johns et al. (2006). This version forms the control over which we measure the improvement in systematic errors of the tropical circulation and summer land surface temperatures and moisture as seen in the final configuration of a new climate version, HadGEM2-AO (Collins et al. (2008)). Analysis of the systematic errors in the climate model are carried out using both atmosphere-only runs (HadGAM1; Table 1) and coupled runs (HadGEM1 and HadGEM2-AO; Table 1). The atmosphere-only runs were forced by prescribed SSTs and sea ice from the second Atmospheric Model Intercomparison Project (AMIP-II; Gates and 15 coauthors (1999)) from 1979 to 1999 and using other boundary forcing as described in Martin et al. (2006). The coupled runs were run in persistent pre-industrial mode using fixed 1860 forcing levels for greenhouse gases, ozone, sulphur, and other aerosol precursor emissions and land surface boundary conditions (see Johns et al. (2006) for further details).

The model results are compared against a range of observations and (re)analyses, depending on the relevant timescale. Generally, the short-range and medium-range predictions

\textsuperscript{1}The Observing system Research and Predictability EXperiment
are compared against either Met Office analyses or radiosonde ascents, while the climate predictions are compared against ECMWF re-analyses (Uppala et al. (2005)). The similarity between the biases in the different configurations compared against the different verification datasets provides confidence that the results are robust.

3. Tropical Performance

a. Analysis of systematic errors on NWP and climate timescales

At NWP timescales there is excessive precipitation and evaporation over tropical oceans. These errors have implications for the water and energy cycles, tropical cyclones and aviation products in the short-range, while on medium ranges, tropical errors can impact extratropical forecasts through teleconnections. As discussed by Johns et al. (2006), the simulation of ENSO is a major weakness of HadGEM1. The climatological trade winds are too strong in the east Pacific, and the associated excessive zonal wind stress in the equatorial region drives excessive upwelling across much of the tropical Pacific. Consequently, HadGEM1 exhibits a marked cold bias in the equatorial Pacific (Johns et al. (2006), their Fig. 3). Johns et al. (2006) showed that the observed eastward shift of the tropical convection during El Niño events, associated with a collapse of the Walker circulation, is not captured in HadGEM1, but that HadGAM1 (with prescribed SSTs) does reproduce the observed features in a more satisfactory manner, suggesting that the problems in HadGEM1 are probably due to the combined interactions of atmosphere and ocean.

The excessive low-level zonal wind is clearly seen in atmosphere-only runs on both climate
and 1-5 day forecast timescales (Fig. 2(b) and (a) respectively). Thus, the source of the wind stress bias in coupled atmosphere-ocean runs of HadGEM1 appears to be the atmosphere model. The zonal wind error is seen to grow steadily in short-range forecasts from analysis time (Fig. 2(a)) and appears to have saturated at day 5 to similar levels (1.5 m s\(^{-1}\)) as seen on climate timescales. This suggests that the error occurs through an immediate response of the model’s physical parametrisations or dynamics and subsequently persists into equilibrium.

As discussed in the introduction, both climate and NWP configurations exhibit similar tropical systematic biases in the thermodynamic fields. We analyse this error growth in further detail by comparing temperature and moisture profiles in MetUM analyses and 36 hour forecasts against radiosonde observations from the ARM Manus Island region (Ackerman and Stokes (2003)). The cold bias at upper tropospheric levels (14-26 km) is already present even in the analysis\(^2\), while the mid-tropospheric warm bias steadily evolves over the first 36 hours of the forecast (Fig. 3). The relative humidity (RH) profiles show a number of interesting discrepancies between analyses, forecasts and radiosondes. Above the freezing level (about 5 km) the analysis is moister than the radiosondes by around 10%, which perhaps could be argued to be within observational uncertainty of measuring RH in moist tropical atmospheres (Ciesielski et al. (2003)). However, the overall structure of decreasing RH with height is consistent with the radiosondes. More interesting is the evolution of the RH in the 30-36 hour forecasts, which moisten by 5-10% relative to the analysis between 15-20 km and dry between 10-15 km. Although some of this RH drift is correlated with the warm-cold bias dipole, similar drying-moistening is also seen in the specific humidity fields themselves (not shown).

\(^2\)The ARM Manus radiosondes were not used in the data assimilation
One hypothesis for this structural change in the humidity profile is that the model’s convective parametrisation is detraining too little moisture in the mid- to upper-troposphere and too much once the convective parcel finally terminates and detrains near the tropopause. The model/analysis hydropause is also too low, suggesting that parametrised convection does not penetrate high enough compared with the radiosondes. This particular error was also noted in the studies of the MetUM convective parametrisation against TOGA-COARE data (Willett et al. (2008)) and in comparisons of modelled tropical brightness temperatures with satellite data (Milton et al. (2001)).

The modelled tropical precipitation on both medium-range (Fig. 4(e,f)) and climate (Fig. 4(c,d)) timescales shows positive biases over the tropical oceans compared with GPCP (Huffman et al. (2001)) and CMAP (Xie and Arkin (1997)) monthly precipitation datasets, with the model biases generally larger than the uncertainty between the observational estimates (Fig. 4(b)). These errors are consistent with excessive diabatic heating of the tropical atmosphere and the growth of the mid-troposphere tropical warm bias discussed above. Although similar on the largest scales, the detailed patterns of error on medium-range and climate timescales are subtly different. This may be related to differences in the initialisation as well as horizontal resolution.

The incorrect distribution of tropical diabatic heating will clearly feed back onto the tropical circulation. For example, idealised modelling experiments by Hartmann et al. (1984) showed strong sensitivity of the Walker circulation and low level winds to the vertical distribution of diabatic heating associated with mature cloud clusters (maximum in upper troposphere) compared with more conventional heating profiles peaking in the mid-troposphere. The 200 hPa velocity potential (Fig. 5) shows excessive divergent outflow over the Indian
Ocean at both climate and medium-range timescales, consistent with the excessive precipitation in these regions (Fig. 4). At climate timescales, there are also two regions of excess subsidence linked to the divergence errors via a north-south Hadley circulation: one over Australia and a second over the east Pacific, while the medium-range forecast shows excess convergence right across the Pacific and into the tropical Atlantic. It appears that, on the medium-range timescale, the errors project more onto east-west Walker circulation than the north-south Hadley circulation. The 200hPa streamfunction analysis fields (Fig. 6(a) and (d)) are dominated by the twin anticyclones of the mean Asian monsoon circulation, straddling the equator with strong easterly flow of the tropical equatorial jet (TEJ) between them. The broad-scale hemispheric biases at both medium-range and climate timescales are similar, with the MetUM showing a tendency to weaken these anticyclones and produce erroneous westerly flow in the Indian Ocean, weakening the TEJ. This excessive upper level westerly flow continues into the Pacific and is consistent with too strong a Walker circulation and too strong easterly flow in the near surface equatorial winds (Fig. 2).

In the following section we discuss some physical improvements to the MetUM convective and boundary layer parametrisations which have an impact on the equatorial wind stress, precipitation, diabatic heating, and overall biases in tropical circulation.

b. Testing proposed solutions - Improvements to physical parametrisations

There has been growing evidence over a number of years that the MetUM convective parametrisation (described in Gregory and Rowntree (1990), Gregory and Allen (1991) and Gregory et al. (1997)) suffered from a number of structural deficiencies highlighted by the
tropical biases in the previous section and discussed in more detail in recent GCSS TOGA-COARE intercomparisons (Petch et al. (2007); Willett et al. (2008)). Further evidence was presented as part of the EUROpean Cloud Systems study (EUROCS) where Derbyshire et al. (2004) showed that mass flux profiles in the MetUM convective parametrisation were unrealistic compared with Cloud Resolving Models (CRMs) and that the convective parametrisation failed to respond correctly to variations in environmental humidity. In response to these deficiencies, a number of changes to the parametrisation have been made. An adaptive detrainment parametrisation for convection has been developed (Derbyshire et al. (submitted)) which relates detrainment to the buoyancy excess of the parcel. This replaces the “forced detrainment” which only detrains when the buoyancy goes below a certain threshold (thus leading to step-changes in the convective updraught mass-flux profile). This results in improved mass-flux profiles in which detrainment occurs more gradually over a greater number of model levels, leading to increased warming in the tropical upper troposphere which reduces the cold bias in this region in the MetUM (Derbyshire et al. (submitted)).

A number of other improvements have also been made to the surface and boundary layer parametrisations. These address known deficiencies in momentum and scalar transports and include (i) modifying stable boundary layer turbulent mixing over the ocean through the use of short-tailed stability functions which better match large eddy simulations (LES), (ii) changes to the surface scalar transfer over the ocean bringing the dependence on wind speed more in line with observations (Edwards (2007)) - the main impacts were a reduction in oceanic latent heat fluxes and reduced precipitation over the oceans which is beneficial in terms of model systematic error (see Fig. 4), and (iii) a non-local scheme for momentum mixing in convective conditions - without these non-local stresses Brown and Grant (1997)
showed that wind profiles in the boundary layer were less well mixed than suggested by LES studies. Changes (i) and (ii) are reported in more detail in Brown et al. (2008). The impacts of changes (i)-(iii) were generally smaller at climate timescales than those seen for adaptive detrainment. For the remainder of this section we will largely focus on the impacts of the adaptive detrainment on tropical performance.

1) IMPACTS ON THERMODYNAMIC FIELDS

Including adaptive detrainment results in smoother mass-flux profiles, which are in better agreement with CRMs (Derbyshire et al. (submitted)). Sensitivity tests carried out in short-range forecasts show the impact on the parametrised convective heating increments (K/day) is twofold (Fig. 7). The shape of the convective heating is changed, with less heating in the mid-tropical troposphere and more heating aloft. The convective heating also penetrates higher in the troposphere, which again reduces a known systematic error in the model (Fig. 3), and is more responsive around the freezing level (about 5km) where we see a minimum in convective heating compared with the old scheme with fixed threshold detrainment. The total (parametrised) diabatic heating also shows similar structural changes arising from convection detraining more gradually and extending higher in the atmosphere.

The impact of these changes in diabatic heating are manifest as reduced model biases in temperature and winds at medium-range and climate timescales (Fig. 8). The days 11-15 cold bias and the equatorial westerly wind bias in the upper tropical troposphere (Fig. 8(c) and (d)) are both reduced compared with the original biases (Fig. 1 (c) and (d)). Similar reductions are also seen in forecast verification against radiosondes ((Derbyshire et al.
At climate timescales the westerly wind biases are also significantly reduced (compare Fig. 8(b) and Fig. 1(b)). The reduction in the temperature biases is discernable (Fig. 8(a) and Fig. 1(a)) but is smaller than that seen in the medium-range forecasts.

2) Impacts on precipitation and tropical circulation

Precipitation is reduced over the tropical oceans on both medium-range and climate timescales when adaptive detrainment is included (Fig. 4(g) and (h)). This is consistent with a reduction in the column-integrated diabatic heating (see Fig. 7). In particular, the excessive oceanic precipitation in the central Indian Ocean is reduced during June-August. This has been a persistent bias in MetUM Asian monsoon predictions at all timescales. Again the impacts are larger at the shorter timescale, where they lead to a significant reduction in the precipitation bias (compare Fig. 4(f) and (j)). The smaller impact at climate timescales may suggest that other longer timescale errors may modulate the adaptive detrainment benefits. This requires further investigation within the seasonal modelling framework currently under development for the next generation HadGEM family of climate models. One area where precipitation biases are worse is in the equatorial east Pacific, particularly at medium-range.

With adaptive detrainment, the divergent errors over the Indian Ocean region are reduced (Fig.s 5(c) and (f)), consistent with the changes in precipitation discussed above. However, on both climate and medium-range timescales, a large divergent error appears in the east Pacific. One possible hypothesis is that, prior to the introduction of adaptive detrainment, the forcing of too strong a Walker circulation and associated strong descent
over the east Pacific tended to suppress convection in that region. The reduced diabatic heating/precipitation in the Indian Ocean and west tropical Pacific with adaptive detrainment improves the Walker circulation (as shown by improved equatorial near surface winds (Fig. 2(b) and (c)), which allows convection to develop more readily in this region. This is perhaps a case where improving the error in one region has removed a compensating error in a remote region.

The introduction of adaptive detrainment and the associated reduction in divergent errors also gives large improvements of around 30% in the rotational flow (Fig. 6(c) and (f)) on a hemispheric scale. In the tropics the excessive westerly flow along the equator at 200hPa is reduced, particularly on climate timescales. In short-range forecast tests the introduction of the non-local momentum mixing (Brown et al. (2008)) also further reduced the excessive wind speeds in the tropical boundary layer, but the remaining boundary layer revisions had little impact on the winds (not shown). The improvements in low level equatorial winds and wind stress have implications for the ENSO response in the coupled model simulations (see section 5b).

4. **Summer land surface error in northern hemisphere continental regions.**

   a. *Analysis of the error in NWP and climate simulations*

   On climate timescales, there are extensive warm biases in daily mean near-surface temperature over northern continents in summer (June to August: JJA) (Fig. 9(a)). Analysis of
daily maximum and minimum temperatures shows that both daytime and nighttime warm biases contribute to this overall error (Fig. 9(b)). These summer warm biases lead to a poor simulation of the boreal forests when this model is coupled to interactive vegetation (Collins et al. (2008)). Major changes in vegetation cover, such as Amazon dieback (Cox et al. (2004)) could have major biogeophysical feedbacks on climate (Betts et al. (2004)). However, reliable predictions of vegetation cover cannot be made if the initial vegetation state in the model contains large systematic errors. On NWP timescales, daytime temperatures over land in boreal summer, which are a key forecast product for customers, are also overestimated (not shown).

Previous studies have highlighted deficiencies in the simulation of clouds and aerosols in the MetUM on climate timescales. Compared with observations, HadGAM1 has too little cloud cover over both Central Asia and North America (Martin et al. (2006)) and similar errors are seen in short-range forecasts (Milton and Earnshaw (2007); Williams and Brooks (2008)). The deficits are mainly in low- and mid-level clouds of thick and intermediate optical depth. These errors in the cloud distribution result in an underestimation of the shortwave cloud radiative forcing (the difference between the total (or all-sky) radiation budget fields and those for cloud-free conditions) over both Eurasia and North America (Martin et al. (2006)). In addition, both shortwave and longwave downward clear-sky fluxes are overestimated in HadGAM1. Aerosol optical depths are underestimated globally in HadGAM1 compared with satellite observations (N. Bellouin, personal communication), and surface measurements (Collins et al. (2008)) and the error in clear-sky radiative fluxes is largely due to the lack of representation of natural (biogenic) continental aerosols and mineral dust aerosols in HadGEM1 (Bodas-Salcedo et al. (2008)).
Examination of the spin-up of the temperature biases over the first few days of an ensemble of HadGAM1 runs reveals that a positive surface temperature bias develops in about 3 days over central Asia, reaching an amplitude of about 2°C (Fig. 10). This warming extends throughout the lower troposphere and is accompanied by a decrease in mid-level cloud amounts and in surface latent heat flux (not shown). This rapid error growth suggests that at least some of the warm bias is related to the model physics rather than changes in atmospheric circulation.

One possible cause of the continental warm bias on climate timescales is that erroneously low summer soil moisture may reduce evaporative cooling of the surface, and hence result in erroneous surface warming. Soil moisture is notoriously difficult to validate as few reliable datasets exist. However, it is possible to compare the soil moisture in HadGAM1 with a previous model version, HadAM3P (Jones et al. (2006)), in which the continental land surface temperatures are more realistic. Over Central Asia, the soil moisture is indeed noticeably lower in HadGAM1 than HadAM3P (Fig. 11). The peak in soil moisture in HadAM3P occurs in April when snowmelt is at its maximum in this region. In contrast, HadGAM1 shows little variation in soil moisture between January and April prior to the summer “dry down”. Further examination reveals that, while the majority of runoff in HadAM3P is subsurface, the surface component is dominant in HadGAM1. This is related to the change in runoff parametrisation from MOSES-I (Cox et al. (1999)) to MOSES-II (Essery et al. (2001)) (see section 4b).

In order to isolate other possible sources of the warm bias we have evaluated boreal summer near surface temperatures, precipitation and fluxes against observations from a site in north-west China (Tongyu) for an 18 day period during July 2003. This site was also
studied in Milton and Earnshaw (2007) but here we extend the comparison to include satellite estimates of cloud and aerosol loading. The site is open grassland with a grass canopy of less than 10cm year round, and although care must be taken in comparing a single site with a model NWP gridbox of 60km, this location represents reasonably homogeneous terrain and vegetation. The observations and their method of measurement are outlined in Table 2.

The air temperature (Fig. 12(a)) shows MetUM has a daytime warm bias of 2-4°C for 10 out of the 18 days, with a mean daily air temperature of 24.1°C compared with 23.5°C for the observations (Table 2). The remaining days have much smaller temperature biases and are characterised either by significant precipitation events, which the model captures reasonably well (Fig. 12(i)), or pristine clear skies such as the 28th and 29th July (low cloud cover, low aerosol loadings, and maximum surface solar insolation (Fig. 12(b),(c) and (d))).

On the non-precipitating cloudy days the downward shortwave (SW) surface radiative fluxes are overestimated by 100-200 Wm$^{-2}$, contributing to the surface warm bias. Possible reasons for this overestimate are (i) lack of cloud cover and/or too small cloud liquid/ice water contents, (ii) underestimates of the column water vapour, or (iii) a lack of aerosol radiative forcing. While it is beyond the scope of this paper to explore each of these hypotheses in depth, we have tried to evaluate model cloud and aerosol against available satellite estimates. Comparison with daily mean cloud fraction estimated from the Moderate Resolution Imaging Spectroradiometer (MODIS) suggests the model underestimates cloud cover (Fig. 12(b)). Comparison of the NOAA daily mean outgoing longwave radiation (OLR) of Liebmann and Smith (1996) with MetUM (Fig. 12(f) and Table 2) shows the day to day variability is actually well captured by the 12-24 hour forecasts, but the model’s tendency is to slightly overestimate OLR on the non-precipitating days, again consistent with too little cloud cover.
or too thin cloud. Comparison of the upward SW radiative fluxes at this site also show an underestimation of the surface albedo which contributes to the warm bias (see Milton and Earnshaw (2007) for discussion).

For the global NWP configurations the aerosol radiative forcing is parametrised by a simple climatology characterising land-sea contrasts in aerosol loading (Cusack et al. (1998)). For the Tongyu site the MetUM aerosol optical depth (AOD) is estimated at a constant 0.25, whereas the MODIS AOD is between 0.5 and 0.8 for most of the period, falling to lower values towards the end of July. The aerosol forcing in this region includes significant quantities of mineral dust blown from the Tibet/Mongolia region (Uno et al. (2006)). Lack of aerosol radiative forcing in the NWP forecasts will clearly contribute to the excessive downward SW flux and warm bias at the surface. The climate configuration already contains a parametrisation of major aerosol species (although there are deficiencies in these schemes as discussed above). There are plans for the NWP configurations to follow this lead.

The surface latent and sensible heat fluxes are both overestimated, with largest errors on the non-precipitating cloudy days, consistent with the downward SW errors (Table 2 and Fig. 12(c), (g) and (h)). The exception is 14-16 and 31 July where latent heat flux is underestimated and sensible heat flux overestimated. This error in Bowen ratio arises from excessively dry soil moisture at this time as shown by comparisons with the soil moisture products from the NOAH and VIC land surface models forced with observed fluxes and precipitation as part of the Global Land Data Assimilation System (GLDAS, Rodell et al. (2004)). This dry bias may be linked to both errors in the treatment of runoff and also to the use of a poor climatological soil moisture used to constrain the initial soil moisture fields. It is possible that, if the soil moisture could be initialised accurately from observations, the
near-surface warm bias may not develop in short-range forecasts. In fact, we may even see a cold bias if the cloud and SW errors were still present and led to excess evaporation. However, the drift in surface temperature would still suggest a physical link between short-range and climate errors. Since August 2005 the NWP configurations have improved initialisation of soil moisture with a nudging scheme using near surface temperatures and humidities (Best and Maisey (2002)) and based on a similar approach to Mahfouf (1991).

In summary, the 12-24 hour NWP forecasts show a similar daytime warm bias to the climate model. Comparison with in-situ and satellite observations suggest the largest errors occur on non-precipitating cloudy days due to lack of cloud or small cloud liquid water contents, resulting in too large downward SW radiation warming of the surface. Clearly the air temperature errors are also affected by (i) the accuracy of the soil moisture contents and partitioning of the SW flux sensible and latent heat fluxes, and (ii) deficiencies in the MetUM surface albedo. The errors in all variables are much smaller on days with either significant precipitation or pristine clear skies (low cloud fractions and AODs).

b. Testing proposed solutions

1) CLOUDS AND AEROSOLS

Following the above analysis, the treatment of clouds and aerosols in the MetUM has been investigated. Examination of the timestep behaviour of convection in the MetUM shows that although instantaneous convective cloud properties are reasonable, the combination of intermittent triggering of convection scheme and the fact that the radiation scheme is only called every 3 hours (for cost reasons) can result in underestimation of the radiative effects
of convectively-generated cloud, including anvil clouds (A. Lock, personal communication). The convection scheme has been modified so that convective cloud properties are allowed to decay exponentially with a 2-hour half-life. This results in more continuous convective cloud and increased average convective cloud amount. The result is some small, but significant, improvements in the northern hemisphere continental temperatures, of around 1°C on climate timescales (Fig. 13(a)) and 0.4°C in short-range forecasts (Fig. 13(d)).

Several changes and additions to the representation of aerosol have been made during the development of HadGEM2-AO (Bellouin et al. (2007)), including a climatology of secondary organic aerosol from biogenic terpene emissions (created using results from STOCHEM (Derwent et al. (2003))) and a mineral dust scheme (Woodward (2001)). These changes improve the agreement in aerosol optical depth between model and observations, allow the seasonal variations in aerosols over the northern hemisphere continental regions to be captured (Bellouin et al. (2007)), and lead to a reduction in the northern hemisphere continental temperatures of between 1 and 2°C (Fig. 13(b)) due to the direct radiative forcing from aerosols. The addition of biogenic aerosol in the NWP configurations at model cycle G44 (Table 1) also contributes to the overall reduction in 1.5m temperature warm bias in forecasts over land during boreal summer (Fig. 13(d)).

2) LAND SURFACE CHARACTERISTICS

Typically, in springtime, large areas of the continental interiors have deep snow cover which melts over initially frozen and saturated soil. The surface parametrisation used in HadAM3P (MOSES-I) deals with the excess water from the snowmelt by adding it to the
downward moisture fluxes into the lower soil layers and eventually into subsurface runoff. However, the scheme used in HadGEM1 (MOSES-II) removes the excess water by adding it straight into the surface runoff if the top soil level is saturated. Thus, a significant proportion of the snowmelt is removed from the terrestrial system during spring, resulting in too little soil moisture by summer. The MOSES-II scheme has been modified in HadGEM2-AO to add such excess water over saturated soil into the downward moisture flux, as in MOSES-I. As a result, the soil as a whole is moister and vegetation suffers less water stress. The change to the treatment of runoff also has a major positive impact on the warm bias, reducing the average surface temperature by up to 4°C over Central Asia and by more than 2°C over North America (Fig. 13(c)). It should be noted that neither approach to the treatment of runoff from saturated soils is ideal. A more realistic approach, whereby some of the excess water percolates downwards and some goes to surface runoff, is preferred and will be worked on in the future.

Following the inclusion of the mineral dust scheme in HadGEM2-AO, corresponding improvements to the albedo have been made, specifically, altering the bare soil albedo to match observations from MODIS (Moody et al. (2005); Houldcroft et al. (2008)) and removing an artificial increase to the Saharan albedo which was made to compensate for the lack of a mineral dust scheme in HadGEM1. These changes have global benefits and also help to alleviate the surface temperature biases over the northern continental regions. Similar improvements to bare soil albedo were implemented in the NWP configuration at cycle G44 (Table 1) with benefits for the surface energy balance and near surface temperatures (Milton et al. (2008)). Further improvements will be made to the albedo of vegetated surfaces following Houldcroft et al. (2008) in future configurations of the MetUM.
c. Impact of combined changes on systematic error

As described in the preceding section, several modifications have been introduced to alleviate the warm bias. When combined, the systematic warm bias over the northern hemisphere continental interiors is reduced substantially on climate timescales (Fig. 9(c)) and in short-range forecasts (Fig. 13(d)), with the surface runoff modification playing the largest role in this change (Fig. 13(c)). In addition to improving the warm bias, the modifications also increase the mean summer precipitation over western Central Asia, where it was too dry previously, and soil moisture increases by up to 25% (not shown). Over North America, the warm bias is also reduced significantly, particularly over the western part of the region. These changes in the mean near-surface air temperature bias are largely achieved through reductions in the maximum daytime temperatures (Fig. 9(d)). This is because several of the changes made to the model physics (e.g. the extended lifetime of convective cloud, the improved representation of aerosols and the changes to surface albedo) mainly affect the daytime conditions. The remaining errors in nighttime minimum temperatures may have several causes and will require further detailed investigation.

5. Overall assessment of improvements

a. Model Metrics

In order to illustrate the beneficial impact of a unified modelling strategy on model development, the performance of the resulting operational configurations of the MetUM at different timescales is now discussed. For many years, an assessment of the overall perfor-
mance of an operational NWP model system has been made using a series of well-defined verification measures (or so-called “model metrics”) for key variables against observations. The WMO-defined standard metrics have proved useful for model development at operational centres to focus on systematic errors and gauge their performance relative to other centres. However, care must be taken in interpreting such measures, given the focus on a relatively small number of variables which can detract attention from improvements in other areas (e.g. model variability and extremes) which may be of increasing importance to customers. The climate modelling community has recently begun to utilise similar sets of measures, to objectively assess both the performance of individual models and changes in the performance of different generations of models, such as those used for different reports of the Intergovernmental Panel on Climate Change (IPCC). Again, these must be applied with a considerable amount of caution, as the value of such simple measures is even less clear when assessing a model for its capability to predict an unknown future climate. Here, we use these metrics but note that for assessing the performance of the MetUM across timescales, we additionally routinely carry out a broader process-based analysis of the models’ capabilities, such as those shown earlier in the paper.

As part of the assessment process of new model/assimilation formulations and as an overall measure of NWP performance the Met Office has used an “NWP Index” for a number of years. The NWP Index is made up of individual skill scores, measured against persistence, for meteorological fields compared with radiosondes, surface observations and model analyses (see appendix A of Rawlins et al. (2007) for details). The revisions to physical parametrisations which improved tropical performance (cycle G39) had a clear beneficial impact on the global NWP index components, with a 3-7% reduction in individual RMS
errors and 2.5 point improvement in the NWP index (Fig. 14). A similar comparison (not shown) for the changes at cycle G44, designed to improve the continental warm bias, showed smaller impacts on the standard NWP index components but clearly had a positive impact on near-surface weather as discussed earlier.

In the climate community, a number of metrics aiming to give an overview of some general assessment of model performance against present day climate observations or reanalyses now exist in the literature (e.g. Murphy et al. (2004); Reichler and Kim (2008a); Gleckler et al. (2008)). Reichler and Kim (2008a) use a composite performance measure ($I^2$) to show that current climate models are more realistic in simulating present-day mean climate than their predecessors. The measure is based on composite normalized mean square errors over a broad range of variables. $I^2$ is derived by first taking differences between simulated and observed mean climate in specific variables and over certain regions. Differences are scaled by the observed interannual variance prior to summing them up, helping to make outcomes from different variables more comparable. The resulting errors are further normalized by the average error found in all models participating in the third Climate Model Intercomparison Project (CMIP-3; PCMDI, 2007), leading to model- and quantity-specific $I^2$ values. Finally, for each model an average $I^2$ is calculated by taking the mean $I^2$ in its individual variables.

Fig. 15 presents quantity-specific $I^2$ values for different regions. The differences between the grey (HadGEM1) and black circles (HadGEM2-AO) demonstrate a clear improvement in HadGEM2-AO against HadGEM1, notably in the tropics (TR) across all of the variables shown here. There is also noticeable improvement in the northern hemisphere (NH) across all of the variables, including surface air temperature over land. The vertical bars in Fig. 15 display the range of outcomes from the 20th century simulations (1979-1999).
of the CMIP-3 models. Although the HadGEM2-AO simulation used pre-industrial forcings, while the CMIP-3 simulations (including HadGEM1) are present-day, Reichler and Kim (2008a) found that the impact of using pre-industrial rather than present-day forcings on the validation against current climate was small compared with the impact of different model generations, and in fact tended to decrease $I^2$. In other words, using pre-industrial simulations for the other CMIP-3 models would widen even more the already existing gap between HADGEM2-AO and the other models. Thus, in terms of simulating present-day mean climate, HadGEM2-AO is overall in a leading position relative to the other CMIP-3 models. This also becomes clear from the final column in each panel, which shows the average $I^2$ across 37 different variables used by Reichler and Kim (2008a,b).

b. Coupled Model performance

1) ENSO

A key motivation for targeting tropical performance in the climate model was to improve the simulation of ENSO over that in HadGEM1. The combined impact of the changes implemented in HadGEM2-AO on the equatorial near-surface winds (Fig. 2(b)) shows a substantial improvement compared with HadGEM1 and this is also seen in the zonal mean wind stress over the critical Niño4 region (Table 3). The changes in surface winds arising from the inclusion of adaptive detrainment are of comparable size to those in HadGEM2-AO (Fig. 2(b)), suggesting that much of the improvement in Niño4 wind stress in HadGEM2-AO arises from this change. Other model changes in HadGEM2-AO, notably changes to the ocean background diffusivity (detailed in Collins et al. (2008)), have also significantly
reduced the mean global SST biases.

Together, we would expect these changes to have a positive impact on the mean state of the equatorial Pacific and the simulation of ENSO in the model, and many improvements compared with observations can be seen in the metrics listed in Table 3. In addition to the significant reduction in bias in mean Niño4 wind stress and Niño3 SST, the surface area of the Indo-Pacific warm pool (a key region for driving global atmospheric circulation) is substantially increased. The amplitude of the SST variability (as measured by the monthly standard deviation of the SST anomaly) across the Niño3 region is improved relative to observations, and composite SST anomalies for El Niño events show an improvement in both the magnitude and spatial extent of the SST anomalies across the Pacific, which leads to a substantially improved response of precipitation to these anomalies (Fig. 16). The maximum precipitation anomaly, which was located to the west of the maritime continent in HadGEM1, has moved eastwards to the west Pacific and a positive anomaly of greater than 0.5 mm day$^{-1}$ is now present over much of the central and eastern equatorial Pacific. The horseshoe pattern of negative anomalies over the maritime continent and the regions extending to the south-east and north-east is also better represented. There is also an improvement in the correlation of Niño3 SST with the Southern Oscillation index (Table 3).

However, the frequency of large El Niño events (>1.5 standard deviation) is reduced in HadGEM2-AO compared with both HadGEM1 and observations (Table 3) and a power spectrum analysis reveals a weak signal at the observed timescale (~4 years), noticeable power at 6-7 years and a dominant peak on decadal timescales (Collins et al. (2008) their Figure 2.5). One of the mechanisms which leads to a change of phase of ENSO from El Niño to La Niña may be related to the “thermocline mode” of variability (Neelin et al.
(1998); Guilyardi et al. (2003); Guilyardi (2006)), which may be weakly simulated in both
HadGEM1 and HadGEM2-AO. This mode seems to be much better simulated in a version of
HadGEM1 with higher ocean and atmosphere resolution, as the ENSO power spectrum and
variability found in that model is much more realistic (Shaffrey et al. (2009)). Due to the
fact that the ENSO in HadGEM1 is confined close to the equator, and may be more of the
“SST mode” form (see Guilyardi (2006)), the ENSO in that model is able to change phase
fairly regularly. In contrast, the ENSO in HadGEM2-AO has a more realistic north-south
extent, but because it lacks the phase-changing process it tends to have a longer timescale
than observed. Hence the removal of one error in the model may reveal other errors which
were previously hidden. Clearly, an improvement in the capability of models to simulate the
change of phase of ENSO will be a target for future models at all resolutions.

2) **Earth System feedbacks**

A primary reason for improving the warm and dry biases in the physical model is to
provide a more realistic surface continental climate for the growth and persistence of char-
acteristic vegetation types when coupled to an interactive vegetation scheme as part of a
full Earth-System model. An indication of whether the package of changes described here
has improved the surface continental climate sufficiently can be gained from examining the
net primary productivity (NPP). This is the difference between the total carbon assimilated
by photosynthesis and the carbon lost through plant respiration. NPP therefore represents
the net uptake of carbon by the vegetation, so it is an important component of the terres-
trial carbon cycle. Although the vegetation distribution in HadGEM2-AO and HadGEM1 is
fixed, we can diagnose the NPP that would arise in a coupled Earth-system model. The impact of the package of changes designed to address the continental near-surface temperature bias is to improve the NPP distribution (Fig. 17) compared with the ISLSCP dataset (the International Satellite Land Surface Climatology Project; Cramer et al. (1999))\(^3\). Whereas HadGEM1 shows significant negative biases in NPP over both continental regions, including some regions where the conditions are unsuitable for any vegetation growth (marked as ‘missing data’ and left blank), with the combined modifications the biases are much smaller. A test of the HadGEM2-ES prototype atmosphere, with these changes included along with the interactive vegetation scheme, shows a substantial improvement in boreal tree density, soil carbon and vegetation productivity, and also in the bare soil distribution, which is a useful prerequisite for improved interactions with the mineral dust scheme.

6. Summary

The reduction of systematic errors in general circulation models is a continuing challenge for improved climate and weather prediction. Feedbacks and compensating errors in climate models often make finding the source of a systematic error difficult. While there may be spin-up errors in short-range forecasts which are not manifest on climate timescales, and similarly errors on climate timescales which only emerge after many months of integration, in this paper we have illustrated how the sources of those systematic biases which appear very early on and persist on long timescales can be identified by the use of the same model across a

\(^3\)The ISLSCP dataset is derived from the mean of 17 terrestrial biogeochemistry models driven by observed climate, and is often used to validate NPP in terrestrial carbon cycle models
range of temporal and spatial scales. Two particular systematic errors have been examined: tropical circulation and precipitation distribution, and summer land surface temperature and moisture biases over northern hemisphere continental regions. Each of these was a cause for concern in both short-range forecasts and in climate simulations. In both cases, the errors were found to develop during the first few days of simulation. The ability to compare in detail the model diagnostics from the first few days of a forecast, initialised from a realistic atmospheric state, directly with observations has allowed deficiencies in the physical parametrisations to be identified which, when corrected, led to improvements on the full range of timescales, from a few days to several decades. There has been a marked improvement in the global NWP index, and the new climate model version, HadGEM2-AO, exhibits enhanced performance over its predecessors, HadGEM1 and HadCM3, and the ensemble of CMIP-3 models.

The unified modelling strategy employed by the Met Office has played a major role in this model development and evaluation process. However, observations of certain key quantities which may have contributed to our analysis of these systematic errors are still lacking across the range of timescales. For example, soil moisture measurements are limited to certain areas of the globe and a reliable long-term global record is lacking (Robock et al. (2000)). This is being addressed through the development of a number of satellite-based measurements such as ASCAT (Bartalis et al. (2007); Wagner et al. (2007)) and the Soil Moisture and Ocean Salinity mission (SMOS; see http://smsc.cnes.fr/SMOS). Similarly, detailed information about cloud processes, ice and water contents and their conversion to precipitation is lacking. Experiments such as CloudSat (Stephens et al. (2002)) are aiming to address this. Rainfall amounts over the oceans are also difficult to estimate due to lack of
in situ observations. Unfortunately, no satellite yet exists which can reliably identify rainfall and accurately estimate the rainfall rate in all circumstances. This is being addressed through measurement projects such as the Tropical Rainfall Measuring Mission (TRMM; see http://trmm.gsfc.nasa.gov) and, in the future, the Global Precipitation Measurement mission (GPM; see http://gpm.gsfc.nasa.gov).

Our study also highlights that the benefits of a unified prediction system across a wide range of timescales may only be fully realised if the full range of timescales is included. Until recently, the operational seasonal and decadal forecasting models used by the Met Office have differed from the model used for weather and climate timescales. This leaves a gap in the range of timescales between the medium-range (15-day) predictions and the centennial climate timescale. The seasonal and decadal modelling frameworks currently under development for the next generation HadGEM family of climate models will enable the development of systematic errors on the monthly to decadal timescale to be studied. Finally, as Hurrell et al. (2009) point out, the drive towards increasingly complex and high resolution models (e.g. those which represent Earth-system feedbacks and regional extreme weather events) emphasises the need for common processes to be addressed in a range of models of different resolution and complexity in order that progress can be made in all.

Acknowledgments.

The authors would like to thank Paul James and Dave Rowell for their initial analyses of the continental surface temperature bias. This work was supported by the Joint DECC and Defra Integrated Climate Programme - DECC/Defra (GA01101).
REFERENCES

*Phys. Today, 56*, 38–44.


Reichler, T. and J. Kim, 2008b: Uncertainties in the climate mean state of global ob-


1 MetUM model configurations

2 Observations and instruments from the Tongyu site. For comparison with the model the data have been averaged from 30 minutes up to 3 hours. The model data are 3-hourly means and instantaneous values at 3-hourly intervals taken from the 12-24 hour NWP forecasts run from both the 00UTC and 12UTC analyses. This allows us to put together eight 3-hourly periods (T+12, 18, 21 & 24 hours from consecutive 00 and 12UTC forecasts) to cover the diurnal cycle.

3 ENSO metrics from HadGEM1 and HadGEM2-AO. Niño3 is defined as the region (5N-5S, 150W-90W) and Niño4 (5N-5S, 160E-150W). IPWP = Surface area of Indo-Pacific Warm Pool (SST ≥ 28 °C). SOI = Southern Oscillation Index. TAUX = surface westerly wind stress.
<table>
<thead>
<tr>
<th>Timescale</th>
<th>Model Configuration</th>
<th>Resolution</th>
<th>Physical formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWP Short range</td>
<td>G32 (May 2004)</td>
<td>N216 (60km) L38</td>
<td>Control (operational JJA 2004).</td>
</tr>
<tr>
<td>(1-5 day)</td>
<td>G38 (Dec 2005)</td>
<td>N320 (40km) L50</td>
<td>Resolution upgrade - increased stratospheric levels</td>
</tr>
<tr>
<td></td>
<td>G39 (Mar 2006)</td>
<td>N320 L50</td>
<td>Revision to model physical parametrizations (i) adaptive detrainment (ii) BL non-local momentum mixing (Brown et al. (2008)), (iii) improvements to oceanic latent heat fluxes (Edwards (2007))</td>
</tr>
<tr>
<td></td>
<td>G44 (May 07)</td>
<td>N320 L50</td>
<td>Revision to physical parametrizations (i) Bare soil albedo based on MODIS (ii) Biogenic aerosols introduced (iii) Changes to runoff (see text) (iv) seasonally varying leaf area index implemented.</td>
</tr>
<tr>
<td>NWP Medium range</td>
<td>MOGREPS-15</td>
<td>N144 (90km) L38</td>
<td>Similar to NWP short-range (see above)</td>
</tr>
<tr>
<td>(1-15 day)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate</td>
<td>HadGEM1</td>
<td>N96 (140km) L38</td>
<td>Coupled model (Johns et al. (2006))</td>
</tr>
<tr>
<td></td>
<td>HadGAM1</td>
<td>N96 (140km) L38</td>
<td>Atmospheric version of HadGEM1 model (Martin et al. (2006))</td>
</tr>
<tr>
<td></td>
<td>HadGEM2-AO</td>
<td>N96 (140km) L38</td>
<td>Coupled model. Revisions to model physical parametrizations, including adaptive detrainment. See Collins et al. (2008).</td>
</tr>
</tbody>
</table>
Table 2. Observations and instruments from the Tongyu site. For comparison with the model the data have been averaged from 30 minutes up to 3 hours. The model data are 3-hourly means and instantaneous values at 3-hourly intervals taken from the 12-24 hour NWP forecasts run from both the 00UTC and 12UTC analyses. This allows us to put together eight 3-hourly periods (T+12, 18, 21 & 24 hours from consecutive 00 and 12UTC forecasts) to cover the diurnal cycle.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model (12-24)</th>
<th>Observation</th>
<th>% error</th>
<th>Obs. details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air Temperature (°C)</td>
<td>24.1</td>
<td>23.5</td>
<td>2.5</td>
<td>1.35m</td>
</tr>
<tr>
<td>SW down surface (W m⁻²)</td>
<td>212.7</td>
<td>187</td>
<td>13.7</td>
<td>Kipp &amp; Zonen pyranometer (CM21) at 2m</td>
</tr>
<tr>
<td>SW up surface (W m⁻²)</td>
<td>36.2</td>
<td>38.2</td>
<td>2.5</td>
<td>Kipp &amp; Zonen pyranometer (CM21) at 2m</td>
</tr>
<tr>
<td>Surface albedo</td>
<td>17%</td>
<td>20%</td>
<td>-15</td>
<td>SWup/SWdown</td>
</tr>
<tr>
<td>LW down surface (W m⁻²)</td>
<td>394.3</td>
<td>399.6</td>
<td>-1.3</td>
<td>Kipp &amp; Zonen pyrgeometer (CG4) at 2m</td>
</tr>
<tr>
<td>LW up surface (W m⁻²)</td>
<td>447.5</td>
<td>445.7</td>
<td>0.4</td>
<td>Kipp &amp; Zonen pyrgeometer (CG4) at 2m</td>
</tr>
<tr>
<td>OLR (W m⁻²)</td>
<td>241.4</td>
<td>238.2</td>
<td>1.3</td>
<td>MODIS daily mean</td>
</tr>
<tr>
<td>Latent Heat (W m⁻²)</td>
<td>85.8</td>
<td>64.8</td>
<td>32.4</td>
<td>Campbell FW05 at 2m</td>
</tr>
<tr>
<td>Sensible Heat (W m⁻²)</td>
<td>33.4</td>
<td>19.9</td>
<td>67.8</td>
<td>Campbell Li-Cor CS7500 at 2m</td>
</tr>
<tr>
<td>Precipitation (mm)</td>
<td>108.89</td>
<td>71.2</td>
<td>53</td>
<td>Texas Elect TE25MM_L</td>
</tr>
</tbody>
</table>
Table 3. ENSO metrics from HadGEM1 and HadGEM2-AO. Niño3 is defined as the region (5N-5S, 150W-90W) and Niño4 (5N-5S, 160E-150W). IPWP = Surface area of Indo-Pacific Warm Pool (SST ≥ 28 °C). SOI = Southern Oscillation Index. TAUX = surface westerly wind stress.

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>HadGEM1</th>
<th>HadGEM2-AO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niño3 standard deviation (°C)</td>
<td>0.8</td>
<td>0.65</td>
<td>0.76</td>
</tr>
<tr>
<td>Average Niño3 1.5σ SST period (years)</td>
<td>8.3</td>
<td>6.9</td>
<td>9.5</td>
</tr>
<tr>
<td>Annual mean Niño3 SST (°C)</td>
<td>25.9</td>
<td>23.7</td>
<td>24.9</td>
</tr>
<tr>
<td>Annual mean Niño4 TAUX (N m⁻²)</td>
<td>-0.029</td>
<td>-0.064</td>
<td>-0.048</td>
</tr>
<tr>
<td>Correlation Niño3 SST/SOI</td>
<td>-0.62</td>
<td>-0.31</td>
<td>-0.45</td>
</tr>
<tr>
<td>Annual mean IPWP (% of obs)</td>
<td>100</td>
<td>13</td>
<td>64</td>
</tr>
</tbody>
</table>
List of Figures

1. (a) Zonally averaged temperature (left column) and zonal wind (right column) biases for June-August: (a) and (b) 10-year mean climatology from HadGAM1 (N96) - ERA40, (c) and (d) days 11-15 of NWP medium range (MOGREPS) forecasts (N144) - Met Office analyses for JJA 2003 and 2006, (e) and (f) days 1-5 of NWP medium-range (MOGREPS) forecasts (N144) - Met Office analyses for JJA 2003 and 2006, (g) and (h) days 1-5 operational NWP forecasts (N216) - Met Office analyses for JJA 2004

2. Equatorial Pacific (5N-5S) 1000hPa zonal wind component for (a) Met Office analyses, day 1, day 3 and day 5 NWP forecasts (cycle G32) during JJA 2004, (b) climatologies from HadGAM1, HadGAM1 & adaptive detrainment (10 year average), HadGEM2-AO (30 year average) for JJA, and (c) impact of revised physics package (cycle G39) vs cycle G38 from trials during August 2005.

3. Mean profiles of (a) temperature error (model-sonde) and (b) relative humidity for July-September 2003 from the MetUM and sondes from ARM TWP site at Manus Island. The sonde data are available at 00UTC and 12UTC. FC0612 = 6 to 12 hour forecast; FC1824: 18 to 24 hour forecast; FC3036: 30 to 36 hour forecast.

47
Tropical Precipitation (40N-40S) in mm/day for (a) CMAP JJA 1979-1998, (b) CMAP-GPCP (v2.0) JJA 1979-98, (c) HadGAM1 JJA 1979-98, (d) HadGAM1-CMAP, (e) medium-range NWP forecasts for JJA 2003 and 2006, (f) medium range NWP forecasts - GPCP, (g) HadGAM1(+ adaptive detrainment) - HadGAM1, (h) Impact of Cycle G39 physics revisions on medium range NWP forecasts (Cycle G39-G38), (i) HadGAM1(+ adaptive detrainment) - CMAP, (j) NWP medium range Cycle G39 - GPCP.

200hPa divergent flow (velocity potential) during June-August (JJA) for (a) ERA-40 analysis (b) HadGAM1-ERA40, (c) HadGAM1(+ adaptive detrainment) - ERA40, (d) Met Oce analysis for JJA 2003 & 2006, (e) Days 11-15 of MOGREPS-15 forecasts JJA 2003 & 2006 minus Met Oce analysis,(f) MOGREPS-15 + cycle G39 physics (includes adaptive detrainment) minus Met Oce analysis. Units are $2^{-6} \text{m}^2\text{s}^{-2}$ for analyses and $1^{-6} \text{m}^2\text{s}^{-2}$ for forecast differences.

As Figure 5 but for 200 hPa rotational flow (streamfunction). Units are $10^{-6} \text{m}^2\text{s}^{-2}$ for analyses and $4^{-6} \text{m}^2\text{s}^{-2}$ for forecast differences.

Impact of revisions to model physical parametrisations at cycle G39 (Table 1) on NWP 1-5 day forecasts showing (a) T+24 convective heating profile at 10N, 60E-90E, and (b) as (a) but for total parametrised heating rate. Units are K/day.
Impacts of adaptive detrainment (& other physics revisions) on zonally averaged temperatures and zonal wind biases during June, July & August. (a) and (b) 10-year mean climatology from HadGAM1 + adaptive detrainment (N96) - ERA40; (c) and (d) days 11-15 of NWP medium range (MOGREPS) forecasts with revised physics (cycle G39 - see Table 1) (N144) - Met Office analyses for JJA 2003 and 2006.

Northern Hemisphere 1.5m temperatures from HadGAM1 compared with a climatology of gridded temperature observations provided by the Climatic Research Unit, Norwich, UK (CRU; New et al. (1999)); (a) HadGAM1 - CRU 1.5m mean temperature, (b) HadGAM1 - CRU 1.5m max. temperature, (c) HadGAM1+revised surface - CRU 1.5m mean temperature, (d) HadGAM1+revised surface - CRU 1.5m max. temperature. Units are K.

Spinup in 1.5m daily mean temperatures over Europe and Asia in an ensemble of 5-day runs of HadGAM1 initialised from ERA-40 re-analysis. (a) Day 2-1; (b) Day 3-1; (c) Day 4-1; (d) Day 5-1. Units are K.

Climatological annual cycle of soil moisture (mm) averaged over Central Asia (60-95E, 45-65N). HadAM3P (solid line), HadGAM1 (dashed line). Figure shows soil moisture (kg m⁻²) over top 3 MetUM sub-surface layers (0-65cm).
Time series of MetUM (12-24 hour forecast) vs observations for 14-31 July 2003 at the GEWEX CEOP/CAMP Tongyu grassland site (44.4N, 122.9E); (a) 1.5m/2m temperatures (3-hourly), (b) Daily mean cloud fraction: MetUM & MODIS, (c) Downward SW surface radiation (3-hourly), (d) Aerosol optical depth: MetUM (climatology; see text) & MODIS (averaged over 40-50N, 118-128E surrounding the CEOP site (see text)), (e) Downward LW surface radiation (3-hourly), (f) OLR from MetUM and NOAA polar orbiting data, (g) Latent Heat Flux (3-hourly), (h) Sensible heat flux (3-hourly), (i) Precipitation accumulation (3-hourly, scaled to mm day$^{-1}$), (j) Soil Moisture content (kg m$^{-2}$) in top 10cm for MetUM and NOAH and VIC land surface models run in GLDAS (see text). The MODIS cloud fraction is a count of cloudy pixels over a 1$^\circ$ box over a day and the MetUM is total cloud fraction (large-scale & convective) in a 0.833$^\circ$ longitude by 0.55$^\circ$ latitude gridbox.

Changes in JJA climatological average surface temperature due to the inclusion of (a) an exponential decay of convective cloud amount, (b) changes to the representation of aerosols, and (c) a change to the treatment of runoff over saturated soil (see text for details). Units are K. Panel (d) shows the 1.5m temperature bias (vs SYNOPS) during JJA 2006 of NWP day 1-5 forecasts run with global model cycle G40 (Control; circles), cycle G42 (squares) showing the impact of introducing convective cloud decay, and cycle G44 (triangles) showing the impact of the full set of physics revisions (see Table 1).

Percentage change in Global NWP index components for MetUM NWP model cycle G39 vs G38 (adaptive detrainment and BL improvements)
Performance measure $I^2$ (Reichler and Kim (2008a)) for different regions and variables, averaged over the four seasons. Lower $I^2$ values indicate better performance, and a value of one indicates the average performance of all CMIP-3 models. Vertical bars show the range of outcomes (best to worst) from the CMIP-3 models. Black circles denote HadGEM2-AO and grey circles HadGEM1. The shown quantities are: precipitation (PR); zonal and meridional wind stress (TAUU, TAUV); surface air temperature over land (TAS); 850 hPa temperature and zonal wind (T850, U850); 200 hPa temperature, zonal winds, streamfunction, and velocity potential (T200, U200, PSI200, CHI200); zonal mean zonal wind (1000-100 hPa) (UA); longwave cloud radiative forcing at TOA (CRFLT); shortwave cloud radiative forcing at TOA (CRFST). AVG indicates $I^2$ averaged across 37 variables, listed in Table 2 of Reichler and Kim (2008b). The shown regions are global (GL), 30-90N (NH), 30S-30N (TR), and 30-90S (SH).

Composite SST (contour; K) and (smoothed) precipitation (shaded; mm day$^{-1}$) anomalies in December-January-February associated with El Niño events: (a) observations (HadISST and CMAP), (b) HadGEM1 and (c) HadGEM2-AO. Observed El Niño events are for the winters of 1982/83, 1987/88, 1991/92 and 1997/98. Model events are defined here for SSTs in the Niño3 region exceeding 1.5 standard deviations of the mean value.
Impact of model improvements on Net Primary Productivity (NPP): (a) HadGEM1 - ISLSCP; (b) [HadGEM1 + surface corrections] - ISLSCP. Values expressed in units of $1.0 \times 10^9$ kg C m$^{-2}$ s$^{-1}$. See text for explanation of ISLSCP dataset.
Fig. 1. (a) Zonally averaged temperature (left column) and zonal wind (right column) biases for June-August: (a) and (b) 10-year mean climatology from HadGAM1 (N96) - ERA40, (c) and (d) days 11-15 of NWP medium range (MOGREPS) forecasts (N144) - Met Office analyses for JJA 2003 and 2006, (e) and (f) days 1-5 of NWP medium-range (MOGREPS) forecasts (N144) - Met Office analyses for JJA 2003 and 2006, (g) and (h) days 1-5 operational NWP forecasts (N216) - Met Office analyses for JJA 2004
Fig. 2. Equatorial Pacific (5N-5S) 1000hPa zonal wind component for (a) Met Office analyses, day 1, day 3 and day 5 NWP forecasts (cycle G32) during JJA 2004, (b) climatologies from HadGAM1, HadGAM1 & adaptive detrainment (10 year average), HadGEM2-AO (30 year average) for JJA, and (c) impact of revised physics package (cycle G39) vs cycle G38 from trials during August 2005.
Fig. 3. Mean profiles of (a) temperature error (model-sonde) and (b) relative humidity for July-September 2003 from the MetUM and sondes from ARM TWP site at Manus Island. The sonde data are available at 00UTC and 12UTC. FC0612 = 6 to 12 hour forecast; FC1824: 18 to 24 hour forecast; FC3036: 30 to 36 hour forecast.
Fig. 4. Tropical Precipitation (40N-40S) in mm/day for (a) CMAP JJA 1979-1998, (b) CMAP-GPCP (v2.0) JJA 1979-98, (c) HadGAM1 JJA 1979-98, (d) HadGAM1-CMAP, (e) medium-range NWP forecasts for JJA 2003 and 2006, (f) medium range NWP forecasts - GPCP, (g) HadGAM1(+ adaptive detrainment) - HadGAM1, (h) Impact of Cycle G39 physics revisions on medium range NWP forecasts (Cycle G39-G38), (i) HadGAM1(+ adaptive detrainment) - CMAP, (j) NWP medium range Cycle G39 - GPCP.
Fig. 5. 200hPa divergent flow (velocity potential) during June-August (JJA) for (a) ERA-40 analysis (b) HadGAM1-ERA40, (c) HadGAM1 (+ adaptive detrainment) - ERA40, (d) Met Office analysis for JJA 2003 & 2006, (e) Days 11-15 of MOGREPS-15 forecasts JJA 2003 & 2006 minus Met Office analysis, (f) MOGREPS-15 + cycle G39 physics (includes adaptive detrainment) minus Met Office analysis. Units are $2^{-6} \text{m}^2\text{s}^{-2}$ for analyses and $1^{-6} \text{m}^2\text{s}^{-2}$ for forecast differences.
Fig. 6. As Figure 5 but for 200 hPa rotational flow (streamfunction). Units are $10^{-6}$ m$^2$s$^{-2}$ for analyses and $4^{-6}$ m$^2$s$^{-2}$ for forecast differences.
Fig. 7. Impact of revisions to model physical parametrisations at cycle G39 (Table 1) on NWP 1-5 day forecasts showing (a) T+24 convective heating profile at 10N, 60E-90E, and (b) as (a) but for total parametrised heating rate. Units are K/day.
Fig. 8. Impacts of adaptive detrainment (& other physics revisions) on zonally averaged temperatures and zonal wind biases during June, July & August. (a) and (b) 10-year mean climatology from HadGAM1 + adaptive detrainment (N96) - ERA40; (c) and (d) days 11-15 of NWP medium range (MOGREPS) forecasts with revised physics (cycle G39 - see Table 1) (N144) - Met Office analyses for JJA 2003 and 2006.
Fig. 9. Northern Hemisphere 1.5m temperatures from HadGAM1 compared with a climatology of gridded temperature observations provided by the Climatic Research Unit, Norwich, UK (CRU; New et al. (1999)); (a) HadGAM1 - CRU 1.5m mean temperature, (b) HadGAM1 - CRU 1.5m max. temperature, (c) HadGAM1+revised surface - CRU 1.5m mean temperature, (d) HadGAM1+revised surface - CRU 1.5m max. temperature. Units are K.
Fig. 10. Spinup in 1.5m daily mean temperatures over Europe and Asia in an ensemble of 5-day runs of HadGAM1 initialised from ERA-40 re-analysis. (a) Day 2-1; (b) Day 3-1; (c) Day 4-1; (d) Day 5-1. Units are K.
Fig. 11. Climatological annual cycle of soil moisture (mm) averaged over Central Asia (60-95E, 45-65N). HadAM3P (solid line), HadGAM1 (dashed line). Figure shows soil moisture (kg m$^{-2}$) over top 3 MetUM sub-surface layers (0-65cm).
Fig. 12. Time series of MetUM (12-24 hour forecast) vs observations for 14-31 July 2003 at the GEWEX CEOP/CAMP Tongyu grassland site (44.4N, 122.9E); (a) 1.5m/2m temperatures (3-hourly), (b) Daily mean cloud fraction: MetUM & MODIS, (c) Downward SW surface radiation (3-hourly), (d) Aerosol optical depth: MetUM (climatology; see text) & MODIS (averaged over 40-50N, 118-128E surrounding the CEOP site (see text)), (e) Downward LW surface radiation (3-hourly), (f) OLR from MetUM and NOAA polar orbiting data, (g) Latent Heat Flux (3-hourly), (h) Sensible heat flux (3-hourly), (i) Precipitation accumulation (3-hourly, scaled to mm day$^{-1}$), (j) Soil Moisture content (kg m$^{-2}$) in top 10cm for MetUM and NOAH and VIC land surface models run in GLDAS (see text). The MODIS cloud fraction is a count of cloudy pixels over a 1° box over a day and the MetUM is total cloud fraction (large-scale & convective) in a 0.833° longitude by 0.55° latitude gridbox.
Fig. 13. Changes in JJA climatological average surface temperature due to the inclusion of (a) an exponential decay of convective cloud amount, (b) changes to the representation of aerosols, and (c) a change to the treatment of runoff over saturated soil (see text for details). Units are K. Panel (d) shows the 1.5m temperature bias (vs SYNOPS) during JJA 2006 of NWP day 1-5 forecasts run with global model cycle G40 (Control; circles), cycle G42 (squares) showing the impact of introducing convective cloud decay, and cycle G44 (triangles) showing the impact of the full set of physics revisions (see Table 1).
Fig. 14. Percentage change in Global NWP index components for MetUM NWP model cycle G39 vs G38 (adaptive detrainment and BL improvements)
Fig. 15. Performance measure $I^2$ (Reichler and Kim (2008a)) for different regions and variables, averaged over the four seasons. Lower $I^2$ values indicate better performance, and a value of one indicates the average performance of all CMIP-3 models. Vertical bars show the range of outcomes (best to worst) from the CMIP-3 models. Black circles denote HadGEM2-AO and grey circles HadGEM1. The shown quantities are: precipitation (PR); zonal and meridional wind stress (TAUU, TAUV); surface air temperature over land (TAS); 850 hPa temperature and zonal wind (T850, U850); 200 hPa temperature, zonal winds, streamfunction, and velocity potential (T200, U200, PSI200, CHI200); zonal mean zonal wind (1000-100 hPa) (UA); longwave cloud radiative forcing at TOA (CRFLT); shortwave cloud radiative forcing at TOA (CRFST). AVG indicates $I^2$ averaged across 37 variables, listed in Table 2 of Reichler and Kim (2008b). The shown regions are global (GL), 30-90N (NH), 30S-30N (TR), and 30-90S (SH).
Fig. 16. Composite SST (contour; K) and (smoothed) precipitation (shaded; mm day$^{-1}$) anomalies in December-January-February associated with El Niño events: (a) observations (HadISST and CMAP), (b) HadGEM1 and (c) HadGEM2-AO. Observed El Niño events are for the winters of 1982/83, 1987/88, 1991/92 and 1997/98. Model events are defined here for SSTs in the Niño3 region exceeding 1.5 standard deviations of the mean value.
Fig. 17. Impact of model improvements on Net Primary Productivity (NPP): (a) HadGEM1 - ISLSCP; (b) [HadGEM1 + surface corrections] - ISLSCP. Values expressed in units of $1.0 \times 10^9$ kg C m$^{-2}$ s$^{-1}$. See text for explanation of ISLSCP dataset.