Response to reviewers: A model-data comparison for a multi-model ensemble of early Eocene atmosphere-ocean simulations: EoMIP

We thank Reviewer #2 and Chris Hollis for their supportive comments.

In addition to the changes made in response to the reviewers (see below), we have: (a) made a number of stylistic and grammatical changes, (b) re-gridded the CCSM ocean data using more standard interpolation code (cdo operators), (c) added further insight about the difference between HadCM and the other models' high-latitude response, and (d) considered *H* as a convergence instead of a divergence in the energy balance analysis.

This response concludes with a 'track changes' version of our manuscript, highlighting clearly the changes made compared with the submitted version [although note that the track changes were produced with latexdiff which has mangled the references, and corrupted the text in a few instances].

We appreciate the detailed consideration given by Chris Hollis to the compilation of proxy data sets used in this model-data comparison. Assembling data from a deep-time interval of earth history is inevitably hampered by sparse data, both geographically and temporally, and the use of disparate proxy methodologies with their own unique uncertainties. The approach taken in the original manuscript was focused on delineating the full range of possible temperature estimates based on the variability inherent in the range of currently available proxy calibrations and methodologies. Inevitably this approach produces relatively wide temperature ranges and may combine a temporal range of data or disparate methodologies that paleoceanographers would not normally countenance. It does, however, robustly determine large-scale proxy-model discrepancies.

Most of the questions raised by Hollis, appear to stem from the implementation of this "full range" approach, including lumping of proxy data across known climatic trends (early Eocene warming) and the reliability of certain proxies from particular geological records (planktic foraminifera oxygen isotopes) or paleogeographic locations (GDGT-based proxies). To answer these valid concerns, we have now attempted to distinguish records from distinct climate states and from different proxy methodologies whilst retaining the *possible* ranges of uncertainty inherent in the underlying data. This revised approach allows for the continued use of these proxy-data comparisons even as our understanding and interpretation of the climate proxy data develops. Below we answer the specific issues raised by Hollis and detail the refinements made to the marine proxy data compilation in answer to these concerns:

1) Time slice refinement: Hollis argues that the model results are specific to particular CO_2 forcing conditions whereas the proxy data – compiled from across the early Eocene – potentially represents a significant range of greenhouse gas forcing. Although robust direct proxy evidence for pCO_2 concentrations across this interval is lacking, there is, as Hollis points out, a well-established warming trend through the early Eocene, culminating in peak-Cenozoic warmth during the early Eocene Climatic Optimum (EECO) (Zachos et al., 2008). This warming trend is most apparent in high latitude sea surface temperature records (Bijl et al., 2009; Hollis et al., 2012) and the global compilation of intermediate-deep water temperatures (Zachos et al., 2008). The concern of Hollis is that the potential comparison of non-EECO and EECO SST estimates contributes to the apparent proxy-proxy data mismatches, especially in the southwest Pacific. It may also contribute to an erroneous assessment of the proxy-derived latitudinal temperature gradient, when high latitude EECO records are

Manuscript prepared for Clim. Past
with version 4.2 of the LaTeX class copernicus.cls.
Date: 19 September 2012

A model-data comparison for a multi-model ensemble of early Eocene atmosphere-ocean simulations: EoMIP

Daniel J. Lunt[1], Tom Dunkley Jones[2], Malte Heinemann[3], Matt Huber[4], Allegra LeGrande[5], Arne Winguth[6], Claire Loptson[1], Jochem Marotzke[7], Chris Roberts[8], Julia Tindall[9], Paul Valdes[1], and Cornelia Winguth[6]

[1]School of Geographical Sciences, University of Bristol, UK
[2]Department of Earth Science and Engineering, Imperial College London, UK
[3]International Pacific Research Center, University of Hawaii, USA
[4]Department of Earth and Atmospheric Sciences, Purdue University, USA
[5]NASA/Goddard Institute for Space Studies, USA
[6]Department of Earth and Environmental Science, University of Texas at Arlington, USA
[7]Max Planck Institute for Meteorology, Hamburg, Germany
[8]The Met Office, Exeter, UK
[9]School of Earth and Environment, University of Leeds, UK

Abstract. The early Eocene (∼55 to 50 Ma) is a time period which has been explored in a large number of modelling and data studies. Here, using an ensemble of previously published model results, making up 'EoMIP' - the Eocene Modelling Intercomparison Project, and syntheses of early Eocene terrestrial and SST temperature data, we present a self-consistent inter-model and model-data comparison. This shows that the previous modelling studies exhibit a very wide inter-model variability, but that at high CO_2, there is good agreement between models and data for this period, particularly if possible seasonal biases in some of the proxies are considered. An energy balance analysis explores the reasons for the differences between the model results, and suggests that differences in surface albedo feedbacks, water vapour and lapse rate feedbacks, and prescribed aerosol loading are the dominant cause for the different results seen in the models, rather than inconsistencies in other prescribed boundary conditions or differences in cloud feedbacks. The CO_2 level which would give optimal early Eocene model-data agreement, based on those models which have carried out simulations with more than one CO_2 level, is in the range 2500 ~~2000~~ ppmv to 6500 ppmv. Given the spread of model results, tighter bounds on proxy estimates of atmospheric CO_2 during this time period will allow a quantitative assessment of the skill of the models at simulating warm climates, which could be used as a metric for weighting future climate predictions.

1 Introduction

Making robust predictions of future climate change is a major challenge, which has environmental, societal, and economic relevance. The numerical models which are used to make these predictions are normally tested over time periods for which there are extensive instrumental records of climate available, typically over the last ∼100 years (**?**). However, the variations in climate over these timescales are small relative to the variations predicted for the next 100 years or more (**?**), and so likely provide only a weak constraint on future predictions. As such, proxy indicators of climate from older time periods are increasingly being used to test models. On the timescale of ∼100,000 years, the Palaeoclimate Modelling Intercomparison project (PMIP, **?**), now in its third phase, is focusing on three main time periods: the mid-Holocene (6,000 years ago, 6k), the Last Glacial Maximum (LGM, 21k), and the Last Interglacial (LIG, 125k). However, these time periods are either colder than modern (LGM), or their warmth is primarily caused not by enhanced greenhouse gases, but by orbital forcing (mid-Holocene, LIG). As such, their use for testing models used for future climate prediction is also limited. On the timescale of millions of years, several time periods show potential for model evaluation, being characterised by substantial warmth which is thought to be driven primarily by enhanced atmospheric CO_2 concentrations. An example is the mid-Pliocene (3 million years ago, 3Ma), when global annual temperature was ∼3°C greater than pre-industrial (**?**). However the latest estimates of mid-Pliocene CO_2 (**??**) range from ∼360 to ∼420 ppmv, which is similar to that of modern (∼390 ppmv in 2010 according to the Scripps CO_2 program, http://scrippsco2.ucsd.edu/), and substantially less than typical IPCC scenarios for CO_2 concentration at the end of this century (∼1,000 ppmv in the A1F1 scenario, >1,370 ppmv CO_2-equivalent in the RCP8.5 scenario, **??**). The time period which shows possibly the most similarity to projections of the end of the 21st century and beyond is the early Eocene, ∼55 to ∼50 Ma. A recent compilation of Cenozoic atmospheric CO_2 is relatively data-sparse during the early Eocene, with large uncertainty range, meaning that values more than 2000 ppmv cannot be ruled out (**?**). Relatively high values for the early Eocene are

Correspondence to: Dan Lunt
(d.j.lunt@bristol.ac.uk)

consistent with recent latest Eocene CO_2 reconstructions of the order 1000 ppmv (**??**). Proxy indicators have been interpreted as showing tropical temperatures at this time $\sim$5 °C warmer than modern (e.g., **?**), and high latitude terrestrial temperatures more than 20 °C warmer (e.g., **?**). Recently, due at least in part to interest associated with this time period as a possible future analogue, there have been a number of new sea surface temperature (SST) and terrestrial temperature data published, using a range of proxy reconstruction methods. There have also been several models recently configured for the early Eocene, and attempts made to understand the mechanisms of Eocene warmth. Most of these studies have carried out some form of model-data comparison; however, the models have not been formally intercompared in a consistent framework, and new data now allows a more robust and extensive evaluation of the models.

The aims of this paper are:

- to present an intercomparison of five models, all recently used to simulate the early Eocene climate.
- to carry out a consistent and comprehensive comparison of the model results with the latest proxy temperature indicators, taking full account of uncertainties in the reconstructions.
- by analysing the energy balance and fluxes in the models, to gain an understanding of the reasons behind the differences in the model results.

Section 2 describes the model simulations, Section 3 presents the datasets used to evaluate the models, and Section 4 presents the model results and model-data comparison. Section 5 quantifies the reasons for the differences between the model results, and Section 6 discusses, concludes, and proposes directions for future research.

2 Model simulation descriptions

Many model simulations have been carried out over the last two decades with the aim of representing the early Eocene. Here, we present and discuss results from a selection of these. We present all simulations of which we are aware that (a) are published in the peer-reviewed literature, and (b) are carried out with fully dynamic atmosphere-ocean General Circulation Models (GCMs), with primitive-equation atmospheres. This makes a total of 4 models - HadCM3L (**?**), ECHAM5/MPI-OM (**?**), CCSM3 (**????**), and GISS ModelE-R (**?**). Criterion (b) is chosen to select the models which are most similar to those used in future climate change projection (i.e. we exclude models with energy balance atmospheres such as GENIE (**?**)). There are two sets of CCSM3 simulations, which we name CCSM_W (**??**) and CCSM_H (**??**). All the models and simulations are summarised in Table 1. Together they make up the 'Eocene Modelling Intercomparison Project', EoMIP. EoMIP differs from more formal model intercomparisons, such as those carried out under the auspices of PMIP, in that the groups have carried out their own experimental design and simulations in isolation, and the comparison is being carried out *post-hoc*, rather than being planned from the outset. As such, the groups have used different palaeogeographical boundary conditions and CO_2 levels to simulate their Eocene climates. This has advantages and disadvantages compared to the more formal approach with a single experimental design: the main disadvantage is that a direct comparison between models is impossible due to even subtle differences in imposed boundary conditions; the main advantage is that in addition to uncertainties in the models themselves, the model ensemble also represents the uncertainties in the paleoenvironmental conditions, and therefore more fully represents the uncertainty in our climatic predictions for that time period.

2.1 HadCM

? investigated the potential role of hydrate destabilisation as a mechanism for the Paleocene-Eocene Thermal Maximum ~~maximum~~ (PETM, $\sim$55Ma), using the HadCM3L model. They found a switch in modelled ocean circulation which occurred between $\times$2 and $\times$4 pre-industrial concentrations of atmospheric CO_2, which resulted in a non-linear warming of intermediate ocean depths. They hypothesised that this could be a triggering mechanism for hydrate release. For the 3 Eocene simulations carried out ($\times$2, $\times$4, and $\times$6), vegetation was set globally to a 'shrub' plant functional type. The paleogeography is proprietary ~~propriety~~ but is illustrated in Supplementary Information of **?**. An additional simulation at $\times$3 CO_2 was carried out with the same model by **?**, which incorporated oxygen isotopes into the hydrological cycle. The $\delta^{18}O$ of seawater from the **?** simulation is used in our SST compilation to inform the uncertainty range of the proxies based on $\delta^{18}O$ measurements (see Section 3).

2.2 ECHAM

? presented an ECHAM5/MPI-OM Eocene simulation and compared it to a pre-industrial simulation, diagnosing the reasons for the Eocene warmth by making use of a simple 1-D energy balance model (which we use in this paper in Section 5). They reported a larger polar warming than many previous studies, which they attributed to local radiative forcing changes, rather than modified poleward heat transport. The Eocene simulation was caried out under $\times$2 CO_2 levels, and a globally homogeneous vegetation was prescribed, with characteristics similar to present-day woody savanna.

2.3 CCSM_W and CCSM_H

? presented a set of Eocene CCSM3 simulations, originally published by **?**, with the main aim of comparing these with a new compilation of proxy terrestrial temperature data. They

found that at high CO_2 ($\times 16$) they obtained good agreement with data from mid and high latitudes. We use this same proxy dataset in this paper, including estimates of uncertainty, for evaluating all the EoMIP simulations.

? and **?** carried out an independent set of CCSM3 simulations motivated by investigating the role of hydrates as a possible cause of the PETM. They found evidence of non-linear ocean warming and enhanced stratification in response to increasing atmospheric CO_2 concentrations, and a shift of deep water formation from northern and southern sources to a predominately southern source.

The CCSM_W and CCSM_H simulations differ mainly in the treatment of aerosols. In the CCSM_W simulation, a high aerosol load is applied, whereas the CCSM_H simulation considers a lower-than-present-day aerosol distribution following the approach by **?**, possibly justified by a reduced ocean productivity and thus reduced DMS emissions. A globally reduced productivity is supported by the recent study of **?**. However, it remains uncertain to which extent intensified volcanism near the PETM might have increased the aerosol load (**?**).

2.4 GISS

? carried out an investigation into the role of the geometry of Arctic gateways in determining Eocene climate with the GISS ModelE-R, configured with $\times 4$ CO_2 and $\times 7$ CH_4 compared with preindustrial. They estimated the change to the total forcing to be about $\times 4.3$ of CO_2 -equivalent, but for the purposes of this paper we assume their simulations were at $\times 4$ CO_2 . They ~~. They~~ found that restricting Arctic gateways led to warming of the North Atlantic and freshening of the Arctic ocean, similar to data associated with the 'Azolla' event (**?**)~~(Brinkhuis et al, 2006)~~. They incorporated oxygen isotopes into the hydrological cycle in their model (**?**), and used the predicted isotopic concentrations of seawater to more directly compare with proxy temperature estimates. The $\delta^{18}O$ of seawater from the **?** simulation is used in our SST compilation to inform the uncertainty range of the proxies based on $\delta^{18}O$

3 Early Eocene SST and land temperature datasets

To evaluate the various climate model simulations, we make use of both terrestrial and marine temperature datasets. The marine dataset has been compiled for this paper, the terrestrial data is identical to that presented in **?**. In both cases we take as full account as possible of the various uncertainties associated with each proxy.

The purpose of these compilations is not to provide a tightly constrained 'time-slice' reconstruction of any point in the early Eocene against which the ensemble, or individual model runs can be compared; instead, we include data spanning the entire early Eocene. This approach is consistent with the EoMIP simulations themselves, in which models have not been run with the same specific set of simulation boundary conditions, such as paleogeography or atmospheric greenhouse gas forcings, but can be considered to reflect a possible range of time periods within the early Eocene.

3.1 Marine dataset

For the purposes of model-data comparison, we ~~We~~ have compiled (see Supplementary data) available paleotemperature estimates for sea surface (GDGT paleothermometry), near-sea surface (mixed layer dwelling planktonic foraminifera) and shallow, inner shelf bottom waters (bivalve oxygen isotopes) ~~,~~ from across the early Eocene (Ypresian stage; ~55.9 to 49Ma).

Long-term paleotemperature records through the early Eocene indicate the presence of ~~This long, ~7 Myr time-span includes~~ a significant warming trend in both oceanic intermediate-waters of ~4 °C (**?**) ~~,~~ and high-latitude sea surface temperatures of up to ~10°C (**??**), in the lead-up to the early Eocene Climatic Optimum (EECO) (**?**). Although ~~, although~~ tropical sea surface temperatures may have been more stable (**?**)

The relative paucity of the available data however, which is taken from a small number of locations, many of which have limited time series and/or poor age control, prohibits a narrowly focused time slice reconstruction of SSTs within the early Eocene. Instead, we have chosen to divide the data into two broad categories, those from the period of peak-Cenozoic warmth during the EECO, and the remainder, assigned to a generally cooler 'background' early Eocene climate state. Pre-PETM records are included in this 'background' early Eocene compilation. Given there is some evidence for warming between pre- and post-PETM conditions in the high latitudes (**??**)

The identification of EECO records is only possible where there is either good age control and/or a clear temperature-trend across a long-term early Eocene record. Only three marine SST proxy data sets are indentified as representing EECO conditions for all or part of the associated SST time series - ODP Site 1172D, Waipara River and the Arctic IODP Site M0004. The EECO in ODP Site 1172D record is indentified following **?** as spanning ~53.1 ~~Also included in the compilation are some data from the very latest Paleocene, within the interval immediately before but not including the Paleocene EoceneThermal Maximum (PETM). These data are included~~ to ~49 Ma (588.85 to 562.70 mbsf), with the pre-EECO interval from 54.9 to 53.3 Ma (611.0 to 591.15 mbsf). All of the Waipara River data used here is identified as representing EECO conditions (**??**). Although the early Eocene age model for the Arctic IODP M0004 is poorly constrained between early Eocene hyperthermal events and the termination of the *Azolla* phase, there is a distinct *cooling* in GDGT-derived proxy temperature estimates between the stratigraphically

lower (core 27X) and upper (cores 23X to 19X) parts of this section (**?**). For the purposes of a more refined proxy-proxy and proxy-model comparison, we have labelled the (warmer) SST data from the lower part of the M0004 as pre-EECO and the (cooler) SST data from the upper part as EECO. This is in line with speculation in **?** that while there is a global ~~increase the geographical coverage of data, especially in the mid to low-latitudes. Given the~~ trend of warming through the early Eocene, Arctic SSTs reduced during this interval. Data from M0004 Core 27X between 369 to 367.9 rmcd has also been excluded, which, based on carbon isotope stratigraphy, likely represents the early Eocene hypethermal event ETM2 (**?**)

The remaining sites have either relatively poorly constrained age models *and* have no discernable SST trend through the dataset used (ODP Sites 690 and 738), are spot-samples within the early Eocene (Tanzania, Hatchetigbee Bluff) or are well-constrained ~~these~~ pre-PETM records. Seymour Island is the exception to this, where there remains uncertainty about even the gross age of this succession. The data used in this compilation is sourced from Telms 3 to 5, which, based on strontium isotope stratigraphy and sparse biostratigraphic data, were thought to span the early Eocene, extending just across the early/middle Eocene boundary (**?**). A revised age assessment, based on dinoflagellate biostratigraphy, suggests that the lower part of this sequence is middle and not early Eocene in age (**?**). This new biostratigraphic data remains somewhat tentative, and while awaiting the publication of a fully revised age model for these successions, the Seymour Island data is provisionally included in the SST compilation. It is however, assigned to the background 'pre-EECO' category, as it appears to be more likely to be representative of middle Eocene post-EECO cooling. It is also noted that although Tanzanian Drilling Project Site 2 (TDP2) was originally reported as extending down into the early Eocene (**??**), with the resolution of planktic foraminifera - nannofossil biostratigraphic mismatches around the early/middle Eocene boundary (**?**), TDP Site 2 is now considered to be entirely within the basal middle Eocene (Paul Pearson, personal communication, July 2012) and is excluded from this compilation~~data points are likely to represent minimum bounds for estimates of early Eocene temperatures~~.

For each location ~~with palaeotemperature estimates,~~ the primary geochemical proxy data were first collated and then used to generate a range of SST ~~set of paleotemperature~~ estimates based on a set of plausible assumptions about the underlying paleotemperature methodology~~the set of calibrations outlined below~~. All of the paleotemperature estimation methods ~~used~~ are subject to ~~a range of~~ uncertainty arising from their present-day calibrations, necessary ~~required~~ assumptions about ancient seawater chemistry and potential non-analogue behaviour between modern and ancient systems. Although positive steps are being made with deep-time proxy inter-comparison studies (**?**), potential non-analogue behaviour ~~the modern and early Paleogene system. Although the latter~~ is very difficult to assess and we do not try to quantify this directly in our uncertainty analysis. We do, however, ~~, we make an~~ attempt to quantify uncertainty associated with both paleotemperature calibrations and the ~~the modern calibrations and~~ estimates of ancient seawater chemistry. This is achieved by 1) applying the standard error determined from the modern calibration data set to paleotemperature estimates; 2) the use of multiple alternate calibrations where there is ongoing debate about the most appropriate calibration or proxy method (GDGT paleothermometry) or where modern calibrations vary with environmental conditions (oxygen isotopes); and 3) apply multiple estimates (Mg/Ca) or different estimation methodologies (oxygen isotopes) for seawater chemistry. Where distinct proxies (GDGT paleothermometry) or distinct parameters for seawater chemistry are used (Mg/Ca and oxygen isotopes) the derived temperature ranges are calculated separately at each site. This leads to the following sets of proxy data: TEX_{86}^{H} , TEX_{86}^{L} , $1/TEX_{86}$, oxygen isotope paleothermometry with modeled and latitudinal corrected $\delta^{18}O_{sw}$

~~, by determining the upper and lower temperature bounds produced by each proxy method across a reasonable set of calibration equations and parameters (outlined below) . To this is added the stated uncertainty determined from the modern calibration dataset.~~

3.1.1 Oxygen isotopes

For planktic ~~planktonic~~ foraminifera-derived $\delta^{18}O$ temperature estimates we applied the temperature-~~calcite~~ $\delta^{18}O$ calibrations of **?** for the symbiotic planktic foraminifera, *Orbulina universa*, under both high and low light conditions (equations 1 ~~temperature calibration for planktonic foraminiferausing both a latitude-corrected estimation of and a location~~ and 2). Together, these two equations bracket most of the natural variability in planktic foraminifera temperature-$\delta^{18}O$ space within modern plankton tow data (**?**). Unlike the multiple GDGT paleotemperature equations, which seem to vary in their accuracy with geographical location or paleoenvironment, the two equations of **?** represent the natural variability at a single location (high/low light conditions). The derived SST estimates from these two equations are thus combined into a single range representing the potential environmental variability at any location. The standard errors on equation 1 (low light) and 2 (high light) are ± 0.7 and $0.5\,^{\circ}C$ respectively.

Three sets of temperature estimate are, however, plotted in Figure 1 for each location with planktic foraminifera $\delta^{18}O$ data based on three estimates of $\delta^{18}O_{sw}$: the latitudinal correction of **?** and the modelled ~~depth-specific (~~mixed-layer $\delta^{18}O_{sw}$ ~~; ~50m depth in the model) modelled estimation~~ of **?** and of **?**. The latitudinal correction is a first-order approximation of the effects of the global hydrological cycle

on seawater $\delta^{18}O_{sw}$ and is a widely used improvement on an 'ice-free' globally uniform estimate of $\delta^{18}O_{sw}$. This empirical relationship does not, however, include zonal deviations from this general pattern. In the early Paleogene world, these zonal deviations are likely to have been significantly different to the modern due to the closure of the major Southern Ocean gateways and the resulting high-latitude isolation of the Atlantic and Pacific basins. The isotope-enabled versions of HadCM3L and GISS, however, reproduce both the expected latitudinal gradients in $\delta^{18}O_{sw}$ and an estimation of early Paleogene basin-to-basin zonal gradients (**??**). These modelled $\delta^{18}O_{sw}$ values, taken from modelled ocean depths of 50m, are a potential refinement to early Paleogene $\delta^{18}O_{sw}$ estimation and are included here to provide a more comprehensive assessment of the range of possible $\delta^{18}O$ SST estimates.

For the latitudinal correction, we assume a global average $\delta^{18}O$ ~~the oxygen isotopic~~ composition of early Eocene seawater of -1‰ to be consistent with the value used by **?**. This is comparable to the -0.96‰ used in a recent early Eocene paleotemperature inter-comparison (**?**), and the value of -0.81‰ used in **?**. A SMOW to VPDB conversion of -0.27‰ was applied to all estimates of $\delta^{18}O_{sw}$ ~~. The published standard error on the calibration is ±1.43~~. The maximum difference in paleotemperature estimates between the three $\delta^{18}O_{sw}$ assumptions is for Seymour Island, where the median value using the modelled $\delta^{18}O_{sw}$ of **?** is 5.6°C cooler than that using the $\delta^{18}O_{sw}$ of **?**.

For the *Eurhomalea* and *Cucullaea* bivalve ~~For the Eurhomalea and Cucullaea bivalve-derived~~ $\delta^{18}O$ data ~~,~~ we used the biogenic aragonite $\delta^{18}O$-temperature calibration of **?** as modified by **?** ~~,~~ with both the latitude-corrected and modelled $\delta^{18}O_{sw}$ noted above. We calculated the standard ~~The published~~ error on the **?** calibration, based on the original data set for biogenic aragonite, to be ~~is~~ ±1.4°C. A SMOW to VPDB correction of -0.2‰ is already implicit within the **?**

3.1.2 Mg/Ca ratios of planktonic foraminifera

To estimate calcification temperature, we used the multi-species sediment trap calibration of **?**, which has a calibration standard deviation of ±1.13°C. This paleotemperature estimation relies strongly upon the assumed value of the Mg/Ca ratio in early Eocene seawater, which is still poorly constrained. Values in ~~There is a considerable disagreement between estimates of Eocene seawater Mg/Ca based on ridge flank hydrothermal carbonate veins, at around 2 mol/mol and the 3 to 4 mol/mol estimates from paired Mg/Ca and oxygen isotope paleothermometry of deep-sea benthic foraminifera (; and see discussion in). Within paleoceanographic studies, it has been typical to use values in~~ the range of 3-4 mol mol^{-1} are typically used within paleoceanographic studies and these produce ~~/mol, which yield plausible~~ tropical (**?**) and mid-latitude (**?**) surface ocean temperatures that are broadly ~~,~~ consistent with independent paleotemperature estimates. ~~There remains, however, a pressing need to understand the causes of this discrepancy and establish robust estimates of the Mg/Ca ratio of ancient seawater.~~ Here, we separately calculate and plot (in Figure 1) paleotemperature estimates based on three values ~~calculate paleotemperatures based on two end members~~ of seawater Mg/Ca across a wide range, namely ~~of~~ 3, ~~and~~ 4 and 5 mol mol^{-1}. Distinguishing temperature estimates based on these three values allows for 1) a clear representation of the sensitivity of temperature to the assumed value of Mg/Ca_{sw}; and, 2) the future use of the most appropriate temperature range if/when more robust constraints on early Paleogene Mg/Ca_{sw} become available.

For reference, an estimate of ~~/mol. This range is based around the estimate of~~ 3.5 mol mol^{-1} is obtained using the **?** calibration for *Oridorsalis umbonatus* and the paired ~~/molcalculated by using their calibration for Oridorsalis umbonatus and values of~~ foraminifera Mg/Ca value of 2.78 mol mol^{-1} and ~~mmol/mol and a~~ $\delta^{18}O$-derived bottom water temperature of 12.4°C they quote for ~~at~~ ~49Ma. A ~~The~~ lower, ~3 mol mol^{-1} ~~/mol~~value is obtained by the same method but using the ~~, but using~~ revised calibrations for *O. umbonatus* (**??**). Higher values of ~~The higher value of~~ 4 to 5 mol mol^{-1} are indicated by $\delta^{18}O$ -Mg/Ca paleotemperature inter-comparisons with well-preserved early Eocene foraminifera (**?**), whilst recent modelling of trace metal fluxes and assessments of the long-term benthic foraminifera $\delta^{18}O$ -Mg/Ca record suggest values of ~3 mol mol^{-1} or less (**??**). These lower values are more consistent with ~~is in line with older~~ estimates based on ridge flank hydrothermal carbonate veins, at around 2 mol mol^{-1} (**?**). There remains a pressing need to understand the causes of these discrepancies and establish robust estimates of the Mg/Ca ratio of ancient seawater (see discussion in **?**)~~attempts to quantify long-term, global trace metal fluxes between the major sources and sinks of Mg and Ca~~ .

3.1.3 TEX_{86}

Determining the appropriate proxy index and calibration for deep-time paleotemperature estimates ~~calibration of the paleotemperature proxy,~~ based on the relative abundances of archaeal-derived ~~several~~ isoprenoid glycerol dibiphytanyl glycerol tetraethers (GDGTs) is problematic~~produced by marine archaeota, to deep-time, warm-climate intervals is an area of active ongoing research~~. Three methods have recently been proposed~~based on the same modern calibration data set~~: separate 'low' and 'high' temperature proxies based on different ratios of GDGTs, ~~and~~ TEX_{86}^{L} and TEX_{86}^{H} (**?**) and a non-linear calibration of the original TEX_{86} index, ~~'~~$1/TEX_{86}$ ~~'~~ (**?**), revised by **?**. TEX_{86}^{H} and $1/TEX_{86}$ ~~the calibrations~~ are based on the same underlying ratio of GDGTs - the original TEX_{86} proxy - but differ ~~differing~~ in the form of their calibration equations (logarithmic versus reciprocal). The fundamentally ~~the~~

~~calibration equation. Whereas and the calibration maintain a consistent relationship of temperature estimates across the range of , the~~ different ratio of GDGT isomers ~~GDGTs~~ within TEX_{86}^{L} results in a proxy that can, in certain instances, produce temperature trends contrary to TEX_{86}^{H} and $1/TEX_{86}$ ~~, means that it canbehave in a fundamentally different manner to these other measures~~. It is, as yet, unclear which of these proxies is the most appropriate for early Eocene paleotemperature estimation. There are indications that their suitability may vary with ~~measure to apply when undertaking deep-time paleotemperature estimation, which may vary across~~ both the temperature range and paleoenvironment of GDGT formation (**?**). ~~(,) range and geographically. New multi-proxy inter-comparisons of these methodologies with oxygen isotope and Mg/Ca paleothermometry suggest that produces the best fit to these independent data in mid- to high-latitude locations (Hollis et al. submitted).~~

For the purposes of this study, we separately calculate and plot (in Figure 1) paleotemperatures at each site ~~attempt to calculate paleotemperature~~ using all three measures: TEX_{86}^{H}, TEX_{86}^{L} and $1/TEX_{86}$. This illustrates the ~~calibration. This provides the~~ full range of temperature estimates produced by these GDGT-based proxies at a given location but also allows for a more refined use of this data as the behaviour of these proxies becomes better understood. The calculation of all three proxies at all sites may be considered by some to be an erroneous application of, for example, a 'low temperature' proxy, ~~all of the most recently proposed proxies and calibrations. In some cases the~~ TEX_{86}^{L} , to the mid and low latitudes of a warm climate state. We must stress that we do not intend to imply that all three are equally applicable at all sites. Rather by showing all three proxies at all sites alongside other proxy temperature estimates, we hope to contribute to the ongoing discussion about the behaviour of GDGT proxies in deep-time paleoenvironments (**?**). Recent development of good practice suggests the exclusion of paleotemperature estimates from samples ~~produces clearly erroneous temperature estimates. These can be limited by the exclusion of all analyses~~ with a BIT index in excess of 0.3 (**?**). Although we accept this as a recommendation, in the existing published data compiled here it would result in the exclusion of all early Eocene data from Tanzania and Hatchetigbee Bluff, which both have BIT indices in the range 0.3 to 0.5. We choose to include this published data, but note these higher BIT indices. In some cases, as for the early Eocene data from Tanzania, TEX_{86}^{L} temperature estimates are clearly erroneous and are excluded. Due to the greater availability of data, samples from ~~, although they do persist in occasional samples from low-latitude locations (Tanzania). We apply a calibration uncertainty of ±2.5to~~ the ~~estimates . temperature estimates from the~~ Arctic Ocean IODP Site M0004 ~~were undertaken on the early Eocene sequence from Core 27X, which is clearly above the PETM interval, to Core 19X . This sequence is below the termination of the Azolla phase in Core 11X which, from correlations with the North Sea, has been assigned to the basal middle Eocene magnetochron C21r . Data within the hyperthermal interval ETM2 and any data points~~ with BIT indices > 0.3 were excluded. The ~~have also been excluded. Throughout this interval the~~ standard TEX_{86} proxies discussed above can be applied to this early Eocene Arctic data rather than the TEX_{86}' proxy used through the PETM by **?**. The errors (°C) on the GDGT-based proxies are ±2.5 for TEX_{86}^{H} (GDGT index-2), ±4.0 for TEX_{86}^{L} (GDGT index-1) and ±5.4 for $1/TEX_{86}$ (**?**).

~~.~~

From the arrays ~~this array~~ of time-varying temperature estimates at each site, and for all assumptions of sewater composition and TEX_{86} calibration, ~~for each site~~we calculated the median, maximum and minimum values from the time series as the basis for the model-data comparisons. There is an important caveat to this approach that relates to the effect of data quantity and stratigraphic range on the temperature envelopes plotted. Where there are reasonably extensive time series, natural temporal variability can result in ~~is data available across much of the early Eocene, stratigraphic/temporal variability leads to~~ a larger envelope of temperature estimates than at sites where data is limited to a few spot samples. As a result~~. Where the data are much more limited in extent~~, these envelopes should ~~are correspondingly smaller. They should thus~~ not be taken to solely represent uncertainty in paleotemperature estimation, but also include a measure of the temporal variability at individual sites~~represent 'error', but instead uncertainty associated with temporal variability~~ .

Finally, in order to provide a single 'zero-order estimate' for SST at each location, we averaged the median estimates at each site across the various assumptions of seawater chemistry and TEX_{86}

3.2 Terrestrial dataset

For the terrestrial, we make use of the data compilation presented in **?**. This is based largely on macrofloral assemblages, with mean annual temperatures being reconstructed primarily by leaf-margin analysis and/or CLAMP (physiognomic analysis of leaf fossils). Other proxies are also incorporated, such as isotopic estimates, organic geochemical indicators, and palynoflora. The error bars associated with each data point incorporate uncertainty in calibration, topography, and dating. More information on the data themselves, and the estimates of uncertainty, can be found in **?**.

Both marine and terrestrial datasets are provided in Supplementary Information, and are plotted geographically in Figures 3 and 4, and latitudinally in Figures 5 and 7.

The SST plots have error bars which include ~~show~~ the contributions from the two sources of uncertainty we have con-

sidered, related to calibration and temporal trends. This approach to the data aims to include a wide range of potential uncertainties in order to highlight both the regions of potential model-data agreement, but more importantly where there appear to be genuine discrepancies that cannot realistically be explained by the uncertainties in the proxy temperature estimations.

4 Results and model-data comparison

In this section, we present results from the EoMIP model ensemble (early Eocene simulations and preindustrial controls), as described in Section 2, and compare them with the data described in Section 3. The reasons for the different model results are explored in more detail in Section 5.

It is useful at this stage to define some nomenclature. To represent the distribution of temperature, we write SST for sea surface temperature (only defined over ocean), or LAT for land near-surface ($\sim$ 1.5m) air temperature (only defined over continents), or GAT for near-surface air temperature (defined globally), or GST for surface temperature (defined globally), or just T for a generic temperature. Global means are denoted by angled brackets, so that e.g. the global mean sea surface temperature is $\langle SST \rangle$. Zonal means are denoted by overbars, so that the zonal mean sea surface temperature is $\overline{SST}$. In the case of model output, ensemble means are denoted by square brackets, such as $[LAT]$. Eocene quantities are given a subscript e, and present/preindustrial (i.e. modern) quantities are given a subscript p. Model values are given a superscript m, and proxy or observed data are given a superscript d. Because the modern observed data has global coverage (albeit interpolated, or assimilated with models in some regions), but the Eocene proxy data is sparse, the modern observed global or zonal means, $\langle T_p^d \rangle$ and $\overline{T_p^d}$ are defined, but the Eocene equivalents are not.

4.1 Inter-model comparison

Figure 2 shows the global annual mean sea surface temperature, $\langle SST \rangle$, and global annual mean near-surface land air temperature, $\langle LAT \rangle$, from all the GCM simulations in the EoMIP ensemble, and for modern observations; the Eocene values are also given in Table 2. The observed modern datasets are HadISST for SSTs (pre-industrial; 1850-1890) and NCEP (**?**) for near-surface air temperatures (present; 1981-2010~~1950-1990~~). For any given CO_2 level, there is a wide range of modelled Eocene global mean values; for example, at 560ppmv, there is a 8.9°C range in $\langle LAT_e^m \rangle$ and a 3.2°C range in $\langle SST_e^m \rangle$. This range is larger than the range of simulated modern global means, which themselves agree well with the observed modern global means. The spread in Eocene results is due to (a) differences in the way the Eocene boundary conditions have been implemented in different models, and (b) different climate sensitivities in the different models. These differences are explored in Section 5. The clustering of the pre-industrial results is likely a result of tuning of the pre-industrial simulations to best match observations. For those models with more than one Eocene simulation, the Eocene climate sensitivity ($\Delta \langle GAT \rangle$ per CO_2 doubling) can also be seen to vary, both between models, and also within one model as a function of CO_2. The variation of climate sensitivity between models is well documented in the context of future climate simulations (e.g. IPCC, 2007). The increase in climate sensitivity with CO_2 (for example in the CCSM_H model) is due to the non-linear behavior of climate system feedbacks, for example associated with water vapour (see Section 5); however, there is also some non-linearity in the forcing itself as CO_2 increases (**?**). For HadCM, it is also related to a switch in ocean circulation which occurs between $\times 2$ and $\times 4$ CO_2 and is associated with a non-linear increase in surface ocean temperature (**?**). The HadCM model also carried out an Eocene simulation with $\times 1$ CO_2 (not shown). Comparison of that simulation with its pre-industrial control shows that changing the non-CO_2 boundary conditions to those of the Eocene (i.e. topographic, bathymetric, vegetation, and solar constant changes) results in a 1.8°C increase in global mean surface air temperature, for comparison with a 3.3°C increase for a CO_2 doubling from $\times 1$ to $\times 2$ under Eocene conditions. At a given CO_2 level, the CCSM_W and CCSM_H models give quite different global means. This difference in mean Eocene climate state between the two similar models is most mostly due to differences in the assumed Eocene atmospheric aerosol loading; CCSM_W includes modern aerosols whereas CCSM_H includes no aerosol loading (see Section 2 and Table 1). Both these models share the same pre-industrial simulation. For all models, the $\langle LAT \rangle$ and $\langle SST \rangle$ means share similar characteristics, albeit with $\langle SST \rangle$ varying over a smaller temperature range.

Figure 3 shows the simulated annual mean SST anomaly from each model, and for the proxy reconstructions. A simple anomaly $SST_e - SST_p$ would not be particularly informative because many regions would be undefined, due to the difference in continental positions between the Eocene and present. Instead, we show $SST_e - \overline{SST_p}$, which is only undefined over Eocene continental regions and latitudes at which there is no ocean in the modern. The figures show that some features of temperature change are simulated consistently across models, such as the greatest ocean warming occurring in the mid-latitudes. This mid-latitude maximum is due to reduced SST warming in the high latitudes due to the presence of seasonal seaice anchoring the temperatures close to 0°C, combined with reduced warming in the tropics due to a lack of snow and seaice albedo feedbacks. However, other patterns are not consistent. For example, GISS at $\times 4$ and HADCM at $\times 6$ have similar values of $\langle SST \rangle$ relative to their controls (8.6 and 9.0 °C respectively), but the warming in GISS is greatest in the northeast Pacific and the Southern Ocean, and the warming in HADCM is greatest in

the north Atlantic and west of Australia. Similarly, ECHAM at ×2 and CCSM_H at ×4 have similar global mean SST anomalies (7.2 and 7.6 ~~7.8~~ °C respectively), but the greatest northern hemisphere warming is in the Atlantic in ECHAM, but in the Pacific in CCSM_H. The two CCSM models exhibit similar patterns of warming, correcting for their offset in absolute Eocene temperature - i.e. the patterns of warming in CCSM_H at ×8 are similar to those in CCSM_W at ×16 (with anomalies of 10.1 and 10.8 ~~10.2 and 10.6~~ °C respectively).

Figure 4 shows the simulated annual mean LAT anomaly from each model, and for the proxy reconstructions. The anomaly is calculated relative to the pre-industrial (or modern in the case of the proxies) global (land plus ocean) zonal mean air temperature for each model, i.e. $LAT_e - \overline{GAT_p}$. The global (as opposed to land-only) zonal mean is used for calculating the anomaly in order to avoid undefined points (for example in the latitudes of the Southern Ocean where there is no land in the modern). Similar to SST, there are some consistent features between models - greatest warming is in the Antarctic (due to the lower topography via the lapse-rate effect and the change in albedo), and there is substantial boreal polar amplification. Again, there are also differences between models. For example, GISS at ×4 and ECHAM at ×2 have similar values of $\langle LAT \rangle$ relative to their controls (8.5 and 7.3 °C respectively), but GISS has a substantially greater warming over southeast Asia. These differences cannot be explained solely by differences in topography - the GISS and ECHAM models both use the Eocene topography of **?**.

4.2 Model-data comparison

Figure 5 shows a zonal SST model-data comparison for each model. The longitudinal locations of the SST data can be seen in Figure 3. Each model is capable of simulating Eocene SSTs which are within the uncertainty estimates of the majority of the data points. The data points which lie furthest from the model simulations are the ACEX TEX_{86} ~~TEX_{86}'~~ Arctic SST estimate (**?**), and the $\delta^{18}O$ and TEX_{86} estimates from the southwest Pacific ~~New Zealand~~ (**?**). The Arctic temperature reconstructions have uncertainty estimates which mean that at high CO_2 ~~(×8-16)~~, the CCSM_H (×8-16) and CCSM_W (×16) model simulations are just within agreement. At this CO_2 level, these models are also consistent with most of the tropical temperature estimates. From Figure 2a, it is likely that other models could also obtain similarly high Arctic temperatures, if they were run at sufficiently high CO_2 or low aerosol forcing. Also, given that some of these models (e.g. HadCM) have a higher climate sensitivity than CCSM_H, this model-data consistency could be potentially obtained at a lower CO_2 than in CCSM~~SUBSCRIPTNBH~~.

TEX_{86} is a relatively new proxy, which, as discussed in section 3, is currently undergoing a process of rapid development. In this context, it has been suggested that the proxy could be recording the palaeotemperature anomaly of the bloom season of the marine archaeota, as opposed to a true annual mean. If this is the case, then it is likely that a more appropriate comparison is with the modelled summer temperature. This is illustrated in Figure 6, for the HadCM and CCSM SUBSCRIPTNBH models~~model~~. In this case, the modelled warm month mean temperature is within within ~~for the highest (×6) is within~~ the uncertainty range of the Arctic TEX_{86} temperatures for both models.

Figure 7 shows the terrestrial temperature model-data comparison for each model. Those models which have been run at high CO_2 (both CCSM models), show good agreement with the data across all latitudes. The other models do not simulate such high temperatures, but, as with SST, it does appear that if they had been run at higher CO_2, the model-data agreement would have been better. The HadCM model appears to be somewhat of an outlier in the northern hemisphere high latitudes, as it shows less polar amplification than the other models (see Section 4.3), an effect also seen in SST.

A quantitative indication of the model-data comparison for each simulation cannot currently be used to rank the models themselves, because the actual CO_2 forcing is not well constrained by data. However, it could give an indication of the range of CO_2 concentrations which are most consistent with the data. Given the sparseness of the SST and terrestrial data, any score should be treated with some caution. This is confounded by the uneven spread of the data; for example, there is a relatively high concentration of terrestrial data in North America. There are also issues associated with the different land-sea masks in the different models, which mean that the number of proxy data locations at which there are defined modelled values differs between the models. Therefore, we generate a simple mean-error score for each simulation, σ, for both SST (σ_{sst}) and land air temperature (σ_{lat}), by averaging the error in temperature anomaly at the location of each of N data points:

$$\sigma_{sst} = \frac{1}{N}\sum(SST_e^m - \overline{SST_p^m} - SST_e^d + \overline{SST_p^d}), \quad (1)$$

$$\sigma_{lat} = \frac{1}{N}\sum(LAT_e^m - \overline{GAT_p^m} - LAT_e^d + \overline{GAT_p^d}), \quad (2)$$

but proceed with caution, being mindful that there is a considerable uncertainty in the score itself. Values of σ for each model simulation are given in Table 2. For each model, the best results are obtained for the highest CO_2 level which was simulated (a result which also applies if an RMS score is used in place of a mean error score). The CCSM_H model at 16× CO_2 has the best (i.e. lowest absolute) values of σ. However, as noted before, it appears that other models would also obtain good σ scores if they had been run at sufficiently high CO_2. A 'best-case' multi-model ensemble can be created by averaging the simulations from each model which have the lowest values of σ (it turns out that those models

with the best σ_{lat} also have the best σ_{sst}). These are the models highlighted in bold in Table 2. The model-data comparison for this multi-model ensemble is shown in Figures 8 and 9. The 2 standard-deviation width of the 'best-case' ensemble overlaps the uncertainty estimates of every terrestrial and ocean proxy data point. However, the high latitude southwest Pacific ~~New Zealand~~ SST estimates are right at the boundary of consistency. The terrestrial data shows very good agreement with the model ensemble, and both data and models show a similar degree of polar amplification (see Section 4.3).

By regressing the CO_2 levels and σ values in Table 2, it is possible (for those models with more than one Eocene simulation) to provide a first-order estimate of the CO_2 level, for each model, which could give the best agreement with the proxy estimates. For HadCM, CCSM_H, and CCSM_W, using σ_{sst} this is 2600ppmv, 4300ppmv, and 4900ppmv ~~2100ppmv, 4100ppmv, and 5400ppmv~~ respectively, and using σ_{lat} this is 2800ppmv, 4500ppmv, and 6300ppmv respectively. These estimates come with many caveats, not least that the uneven and sparse data spread means that the absolute minimum mean error, σ, is not necessarily a good indicator of the correct global mean temperature. However, they do indicate the magnitude of the range of CO_2 values that could be considered consistent with model results. These values are significantly higher than those presented for this time period in the compilation of **?**.

4.3 Meridional gradients and polar amplification

The changes in meridional temperature gradient are summarised in Figure 10, which shows the surface temperature difference between the low latitudes ($|\phi| < 30°$) and the high latitudes ($|\phi| > 60°$) as a function of global mean temperature, and how this is partitioned between land and ocean warming (Figure 10b, ϕ is latitude in degrees). All the Eocene simulations have a reduced meridional surface temperature gradient compared with the pre-industrial, and the gradient reduces further as CO_2 increases, i.e. polar amplification increases (Figure 10a). However, there is a high degree of inter-model variability in the absolute Eocene gradient, with HadCM appearing to be an outlier with a relatively high Eocene gradient. There is some indication that the models are asymptoting towards a minimum gradient of about 20°C. This, along with our energy flux analysis (see Section 5), supports previous work (**?**) ~~(Huber et al, 2003)~~ that implied that meridional temperature gradients of the order 20°C were physically realistic, even without large changes to meridional heat transport. Compared with preindustrial, the meridional surface temperature gradient reduces more on land than over ocean (Figure 10b). For HadCM, this applies also to the Eocene simulations as CO_2 increases. However, for the two CCSM models, the meridional temperature gradient is reduced by a similar amount over land and ocean as a function of CO_2, with some indication, at maximum ($\times 16$) CO_2, that the SST gradient starts reducing more over ocean than over land. This implies that when considering changes relative to the modern, it is possible to have substantially different temperature changes over land compared with over ocean at the same latitude. This is also clear from comparing Figure 3 with Figure 4, and shows the importance of differentiating terrestrial and oceanic signals when considering the consistency between different proxy data, and between data and models.

5 Reasons for inter-model variability: an energy flux analysis

It is interesting up to a point to simply intercompare model results, and to compare with data, but also of interest is to know *why* different models behave differently. Given the complexity of climate models, this can be problematic, and traditionally, groups such as PMIP have not often diagnosed in detail the differences. Here, we attempt to diagnose some aspects of the differences between the model results, building on a 1-D energy-balance approach as outlined by **?**. Here, the causes of the zonal-mean temperature response of a model are diagnosed from the top-of-the-atmosphere and surface radiative fluxes, including their clear-sky values, assuming simple energy balance. Any difference between the meridional temperature profile in the GCM, and that estimated from the energy-balance approach, is attributed to meridional heat transport. As such, the change in meridional temperature profile between two simulations (such as a pre-industrial control and an Eocene simulation) can be attributed to a combination of (1) changes in emissivity due to changes in clouds, (2) changes in emissivity due to changes in the greenhouse effect (i.e. CO_2 and water vapour concentration changes, and lapse-rate effects), (3) changes in albedo due to changes in clouds, (4) changes in albedo due to Earth-surface and atmospheric aerosol changes, and (5) changes in meridional heat transport.

Following **?**, the 1-D energy balance model (EBM) is formulated by equating the incoming solar radiation with outgoing long wave radition, with any local inbalance attributed to local meridional heat transport:

$$SW_t^{\downarrow}(1-\alpha) + H - H = \epsilon\sigma\tau^4 \quad (3)$$

where $SW_t^{\downarrow}$ is the incoming solar radiation at the top of the atmosphere, α is the planetary albedo, H is the net meridional heat transport convergence~~divergence~~, ϵ is the atmospheric emissivity, σ is the Stephan-Boltzmann constant, and τ is the surface temperature, to be diagnosed by the EBM. All variables are functions of latitude apart from σ.

The planetary albedo is given by

$$\alpha = \frac{SW_t^{\uparrow}}{SW_t^{\downarrow}} \quad (4)$$

and the atmospheric emissivity is given by

$$\epsilon = \frac{LW_t^{\uparrow}}{LW_s^{\uparrow}}, \quad (5)$$

where $SW_t^{\uparrow}$ and $SW_t^{\downarrow}$ are the outgoing and incoming top of the atmosphere shortwave radition respectively, and $LW_t^{\uparrow}$ and $LW_s^{\uparrow}$ are the upwelling top of the atmosphere and surface longwave radiation respectively. Given that the surface emits long wave radition according to

$$LW_s^{\uparrow} = \sigma\tau^4, \quad (6)$$

it follows that the meridional heat transport convergencedivergence, H, is given by

$$H = -+(SW_t^{net} + LW_t^{net}),. \quad (7)$$

where the superscript net represents net flux (positive downwards). Equation 7 reflects the necessity that, in equilibrium, any net downward shortwave plus longwave heat flux has to be compensated by a negative meridional heat flux convergence (note that there is a typo N.B. the '+' sign in this equation, which was wrongly given as '-' in the equivalent equation of in ?, the minus sign in their Equation 4 is erroneous). All the radiative fluxes are output directly from the GCMs, and used as input into the energy balance model. From Equation 3, it follows that

$$\tau = (\frac{1}{\epsilon\sigma}(SW_t^{\downarrow}(1-\alpha)+H)^{0.25})-H)^{0.25} \equiv E(\epsilon,\alpha,H) \quad (8)$$

The difference in temperature between two simulations, $\Delta T = \tau - \tau'$ is given by $E(\epsilon,\alpha,H)$-$E(\epsilon',\alpha',H')$, where the prime, $'$, represents values in the second simulation. In order to diagnose the reasons for the temperature differences in two simulations, we consider changes to the diagnosed emissivity, planetary albedo, and heat transport, and write,

$$\Delta T_{emm} = E(\epsilon,\alpha,H) - E(\epsilon',\alpha,H) \quad (9)$$

$$\Delta T_{alb} = E(\epsilon,\alpha,H) - E(\epsilon,\alpha',H) \quad (10)$$

$$\Delta T_{tran} = E(\epsilon,\alpha,H) - E(\epsilon,\alpha,H'), \quad (11)$$

where ΔT_{emm}, ΔT_{alb}, and ΔT_{tran} are the components of ΔT due to emmissivity, planetary albedo, and heat transport changes respectively. Because the changes in emissivity, albedo, and heat transport are relatively small compared to their magnitude,

$$\Delta T \simeq \Delta T_{emm} + \Delta T_{alb} + \Delta T_{tran}. \quad (12)$$

We further partition the ΔT_{emm} and ΔT_{alb} terms by considering the clear-sky radiative fluxes, also output directly from the GCMs. Using cs as a subscript to denote clear-sky fluxes, we can estimate the contribution due to the greenhouse effect (CO_2 and water vapour and laspe rate) changes, ΔT_{gg}, and the contribution due to surface albedo and aerosol changes, ΔT_{salb},

$$\Delta T_{gg} = E(\epsilon_{cs},\alpha_{cs},H_{cs}) - E(\epsilon'_{cs},\alpha_{cs},H_{cs}) \quad (13)$$

$$\Delta T_{salb} = E(\epsilon_{cs},\alpha_{cs},H_{cs}) - E(\epsilon_{cs},\alpha'_{cs},H_{cs}) \quad (14)$$

because the emmissivity change in the clear-sky case is solely due to greenhouse effect changes, and the albedo change in the clear-sky case is mainly due to surface albedo and aerosols. Considering the remaining temperature difference as due to clouds, we can then write

$$\Delta T_{lwc} = \Delta T_{emm} - \Delta T_{gg} \quad (15)$$

$$\Delta T_{swc} = \Delta T_{alb} - \Delta T_{salb}, \quad (16)$$

where ΔT_{lwc} and ΔT_{swc} are the components of ΔT due to long-wave cloud changes and short-wave cloud changes respectively. In this way, a temperature difference between two simulations can be partitioned into 5 components, given by Equations 9-11 and 13-16.

Figure 11 shows the results from this energy balance analysis, for a number of pairs of simulations. Figures 11(a-c) show the models which simulate a transition from pre-industrial to Eocene at $\times 2$ CO_2. ECHAM and CCSM_H show similar results in terms of the reasons for this change. They show a high latitude warming in both hemispheres caused mainly by non-cloud albedo changes, with a significant contribution also from emissivity changes. In both these models, short-wave cloud albedo changes act to reduce the polar amplification in both hemispheres. The greater global temperature change in ECHAM compared with CCSM_H is due to the greater change in greenhouse effect. However, the energy balance analysis does not allow us to diagnose if this is due to a greater radiative forcing given the same CO_2 increase, or due to greater water vapour feedbacks or lapse-rate changes in ECHAM. HadCM exhibits quite different behavior. In the Southern Hemisphere, the zonal mean temperature increase is due predominantly to non-cloud albedo changes, and is reduced relative to the other two models. In the Northern Hemisphere, the increase in temperature is much reduced relative to the other two models, due to a lack of non-cloud albedo feedbacks, and changes in emissivity. ? Abbot and Zziperman (2008) suggested that the lack of sea ice in the Arctic can lead to stronger convection over the relatively warm Arctic sea surface during winter, leading to more convective clouds and increased water vapour concentrations, and thereby causing polar amplification via both albedo and emissivity effects. The largely decreased (versus unchanged) surface albedo in northern high latitudes in CCSM_H and ECHAM (versus HadCM), increased (versus virtually unchanged) longwave cloud radiative forcing, and reduced (versus hardly changed) clear-sky emissivity indicates that this seaice/convection feedback is active for $\times 1$ to $\times 2$ in CCSM_H and ECHAM, but absent in HadCM. The reduced strength of this feedback in HadCM may be related to the relatively strong Eocene seasonality

in HadCM compared with the other models (as can be seen by comparing Figure 6 with Figure 5), which suppresses Arctic convection in HadCM in winter. Changes in heat transport are playing a relatively minor role in determining the latitudinal temperature profile in ECHAM, CCSM SUBSCRIPTNBH, and HadCM, which supports previous findings using ECHAM alone (**?**)

Figures 11(d-g) show the models which simulate a transition from pre-industrial to Eocene at $\times 4$ CO_2. For HadCM and CCSM_H, the results are very similar to at $\times 2$ CO_2, but with greater magnitude; for both models each component contributes the same fraction to the total warming under $\times 2$ as to under $\times 4$, to within ~10%. CCSM_W is very similar to CCSM_H, except that it has reduced warming due to decreased change in non-cloud albedo. This is most likely a direct result of the different aerosol fields applied in the these two models for the Eocene (see Table 1). The model which exhibits the greatest warming is the GISS model. This high sensitivity relative to the other models is due to greater greenhouse gas effect changes~~forcing~~, and greater cloud albedo feedbacks. The warming over Antarctica is particularly large in the GISS model, and is due to a greater local change in non-cloud albedo. However, the GISS model also has strong negative cloud forcing at high latitudes in both hemispheres.

Figures 11(h-i) show the models which simulate a transition from $\times 2$ to $\times 4$ CO_2 under Eocene conditions. HadCM has a greater climate sensitivity that CCSM_H, and this is due to greater changes in greenhouse gas emissivity, and a positive as opposed to negative cloud albedo feedback. The relative lack of polar amplification in both models compared to the results discussed above, is due to the lack of Antarctic ice sheet in the Eocene. The small amount of polar amplification in HadCM is due to changes in heat transport; in CCSM it is due to non-cloud albedo changes in the Northern Hemisphere.

Figures 11(j-k) show the models which simulate a transition from $\times 4$ to $\times 8$ CO_2 under Eocene conditions. Similar to the transition from $\times 2$ to $\times 4$, the polar amplification is relatively small. The warming is due almost entirely to the changes in emissivity (direct CO_2 forcing and water vapour feedbacks and laspe-rate changes), and unsuprisingly has a similar latitudinal distribution in the two models. However, in the Northern Hemisphere high latitudes the CCSM_H model shows strong opposing effects of cloud and surface changes, which are not present in CCSM_W. This is most likely due to the remnants of Arctic seaice in CCSM_W at $\times 8$ CO_2 which are not present in the warmer CCSM_H model. Comparison of Figure 11(k) with Figure 11(i) shows that the increase in climate sensitivity in CCSM_H as a function of background CO_2 is due almost entirely to increased non-cloud emissivity changes; the framework does not allow us to determine if this is due to increasing radiative effects due to CO_2, or increasing water vapour feedbacks or laspe-rate changes. However, it is clear that it is not due to increased albedo feedbacks, or cloud processes.

Given that the models prescribe Eocene vegetation in quite different ways, it is interesting to assess how much this affects inter-model variability. Figure 12 shows the surface albedo in the pre-industrial control and the $\times 2$ CO_2 simulations for HadCM, CCSM_H, and ECHAM. At the high latitudes, this is affected by snow and sea ice cover and prescribed changes in ice sheets, but at low latitudes this is purely a result of the imposed vegetation and open-ocean albedos. The fact that all the models have a low latitude albedo which is similar to their control, and similar to each other, indicates that this aspect of experimental design is likely not playing an important role in determining the differences in results between the models.

6 Conclusions and Outlook

We have carried out an intercomparison and model-data comparison of the results from 5 early Eocene modelling studies, using 4 different climate models. The model results show a large spread in global mean temperatures, for example a $\sim 9^{\circ}$C range in surface air temperature under a single CO_2 value, and are characterised by warming in different regions. The models which have been run at sufficiently high CO_2 show very good agreement with the terrestrial data. The comparison with SST data is also good, but the model and data uncertainty only just overlap for the Arctic and southwest Pacific ~~New Zealand~~ $\delta^{18}O$ and TEX_{86} proxies. However, if a possible seasonality bias in the proxies is taken into account, then the model data agreement improves further. We have interrogated the reasons for the differences between the models, and found differences in climate sensitivity to be due primarily to a combination of greenhouse effect and surface albedo feedbacks, rather than differences in heat transport or cloud feedbacks.

There are several issues which have emerged from this study, which should be addressed in future work aimed at reconciling model simulations and proxy data reconstructions of the Early Eocene (many of which also apply to other time periods).

Firstly, modelling groups should aim to carry out simulations over a wider range of atmospheric CO_2 levels. In particular, the results of CCSM_H indicate that at high prescribed atmospheric CO_2 and low aerosol forcing, the models and data come close together. Some of this work is in progress (e.g. simulations at $\times 3$ CO_2 are currently being analysed for the ECHAM model). However, it should be noted that this is not always possible. For example, the Eocene HadCM model has been run at $\times 8$ CO_2, but after about 2700 years the model developed a runaway greenhouse, and the model eventually crashed (**?**). A similar effect has been observed in the ECHAM model at $\times 4$ CO_2 (**?**). Whether such an effect is 'real', i.e. whether the real world would also develop a runaway greenhouse, is completely unknown. In any case, it is clear that modelling the early Eocene climate pushes

the climate model parameterisations to the boundaries within which they were designed to operate, if not beyond these boundaries.

Some of the differences between the model results can be attributed to differences in the experimental design. In particular, some models apply a very generic Eocene vegetation, which is not particularly realistic. A slightly more coordinated study could provide guidelines for ways to better represent Eocene vegetation, for example by making use of palynological data, or by using dynamic vegetation models where available. This would provide an ensemble of model results which better represented the true uncertainty in our model simulations. Other inconsistencies between model simulations should not necessarily be eliminated - for example, different models using different paleogeographical reconstructions may be more representative of the true spread of model results than if all groups used a single paleogeography.

On the data side, better understanding of the temperature proxies and their associated uncertainties, in particular seasonal effects, is a clear goal for future work, as is greater geographical and finer temporal coverage.

Perhaps most crucial of all, better CO_2 constraints from proxies would be of huge benefit to model-data comparison exercises such as this. Recently, much work is being undertaken in this area, but this should be intensified wherever possible. We note that at high CO_2, due to the logarithmic nature of the CO_2 forcing, proxies which may have relatively coarse precision at low CO_2, can actually provide very strong constraints on the CO_2 forcing itself. Such constraints on CO_2, combined with proxy temperature reconstructions with well defined uncertainty ranges, could provide a strong constraint on model simulations, providing quantitative metrics for assessing model performance, and could ultimately provide relative weightings for model simulations of future climates.

Acknowledgements. This paper resulted from discussions at the 'CO_2 at the zoo' event held in Bristol in 2010, and from a Royal Society Kavli meeting held in 2011. DJL acknowledges funding as an RCUK research fellow. NCEP Reanalysis Derived data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.esrl.noaa.gov/psd/.

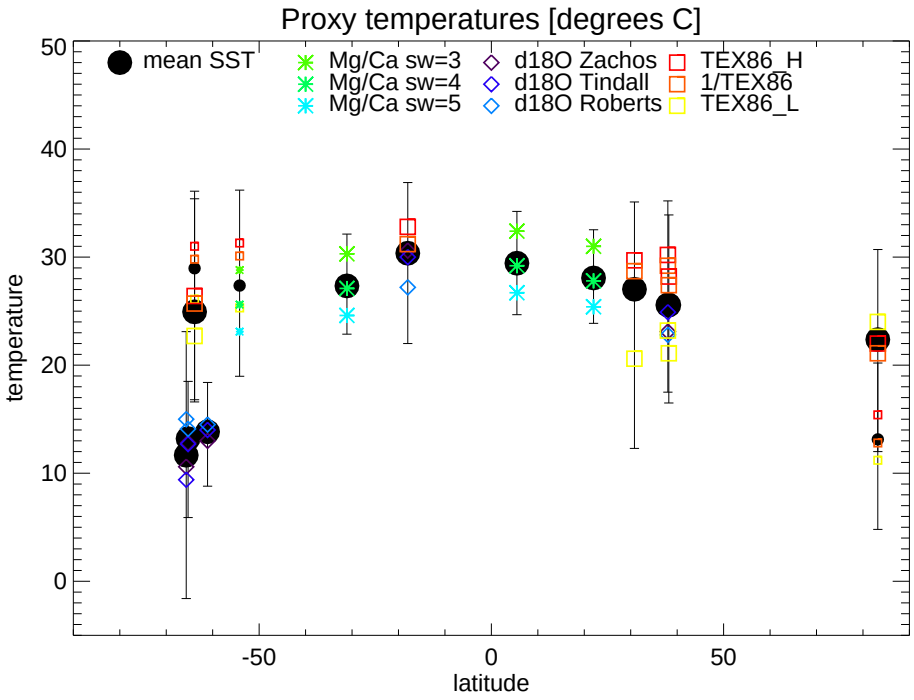
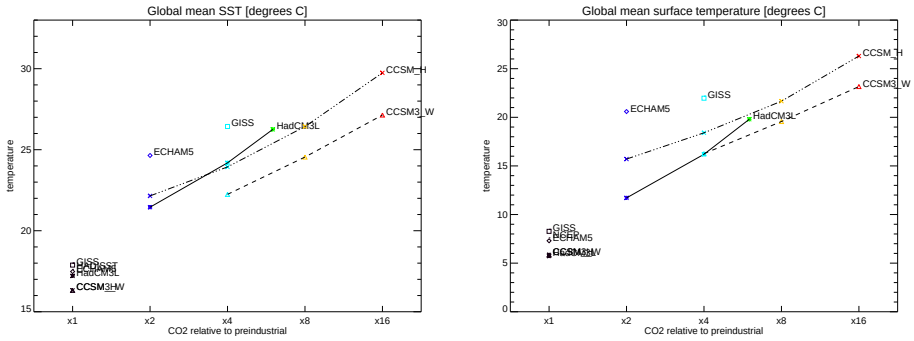


Fig. 1. SST proxy dataset, compiled for this study. Coloured symbols show the various median estimates from the literature, with various asumptions about Mg/Ca of sewater, $\delta^{18}O_{sw}$, and TEX_{86} calibration. Error bars indicate the maximum and minimum range at each site including temporal variability and calibration error. The filled black circles represent the mean SST at each site, averaged over the various assumptions. Larger symbols represent 'background' early Eocene state, smaller symbols represent the EECO. See Section 3.1 for more details.



(a) (b)

Fig. 2. Global annual mean (a) SST ($\langle SST \rangle$) and (b) continental 2m air temperature ($\langle LAT \rangle$), as a function of CO_2 for all simulations, and for observational datasets. The simulations at $\times 1$ CO_2 are pre-industrial reference simulations.

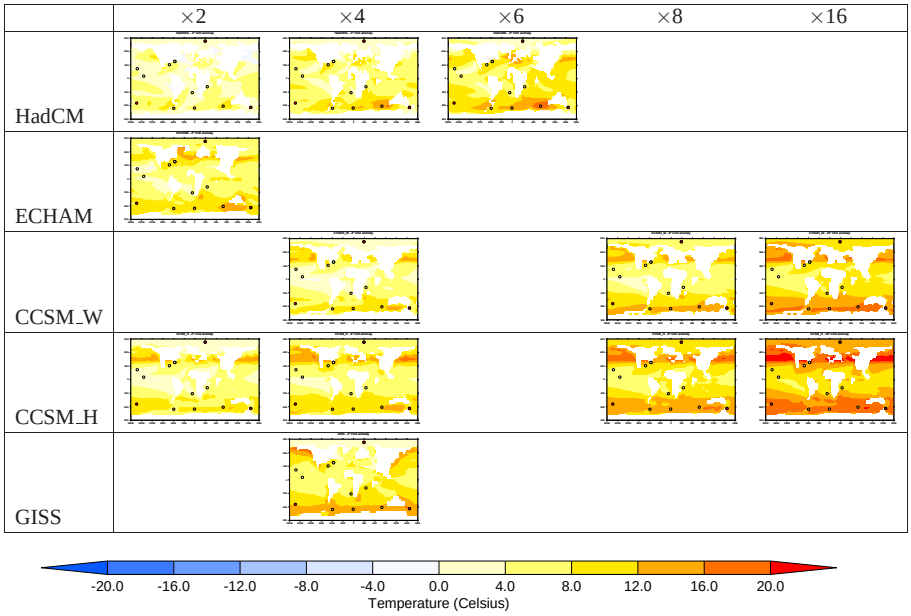


Fig. 3. SST anomaly in the model simulations ($SST_e^m - \overline{SST_p^m}$), as a function of model and fractional CO_2 increase from pre-industrial. Also shown for the proxies are $SST_e^d - \overline{SST_p^d}$.

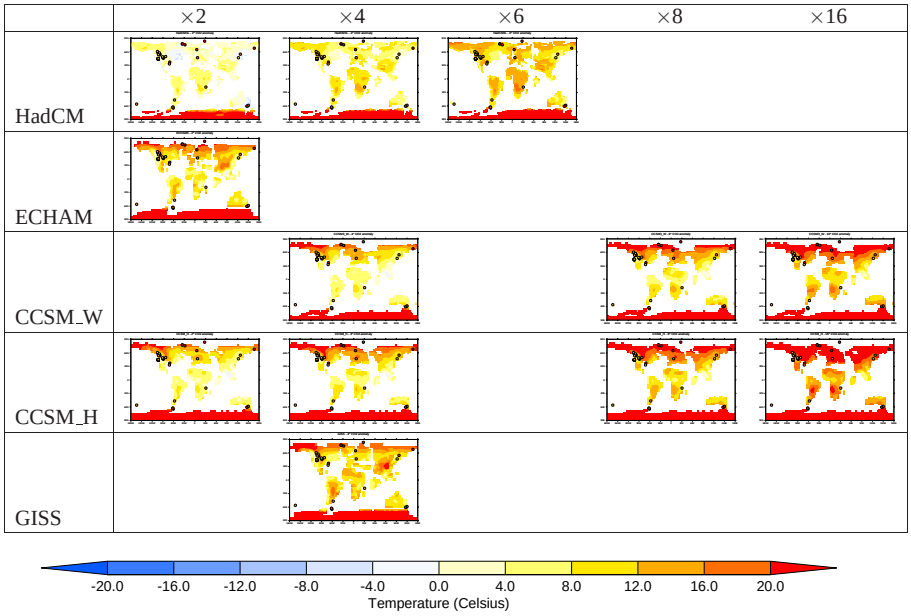


Fig. 4. Continental surface air temperature anomaly in the model simulations ($LAT_e^m - \overline{GAT_p^m}$), as a function of model and fractional CO_2 increase from pre-industrial. Also shown for the proxies are $LAT_e^d - \overline{GAT_p^d}$.

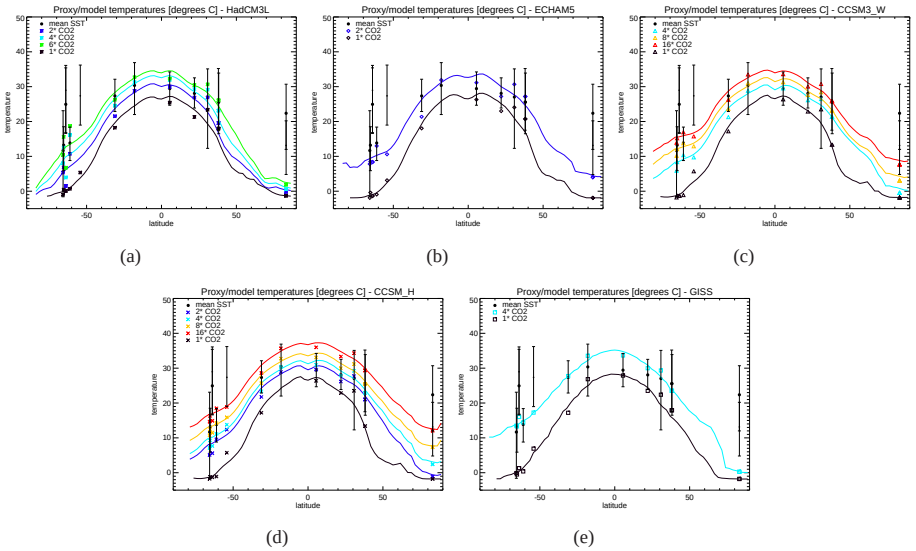


Fig. 5. Comparison of modelled SST with proxy-derived temperatures, SST vs. latitude. The simulations at $\times 1$ CO_2 are pre-industrial reference simulations. For the model results, the continous lines represent the zonal mean, and the open symbols represent the modelled temperature at the same location (longitude,latitude) as the proxy data. For the proxy data, the filled symbols represent the mean proxy temperature, and the error bars represent the range. The smaller filled symbols are EECO temperatures~~range is made up of two components: calibration uncertainty (black bar) and temporal uncertainty (grey bar)~~. See Section 3 and Supplementary Information for more details of the range calculations.

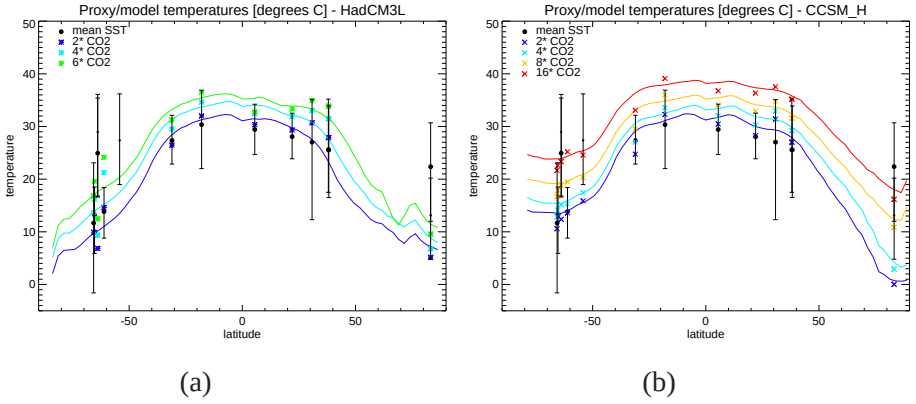


Fig. 6. As Figure 5a and d, but the HadCM and CCSM SUBSCRIPTNBH modelled zonal mean represents the warm month mean SST, as opposed to annual mean.

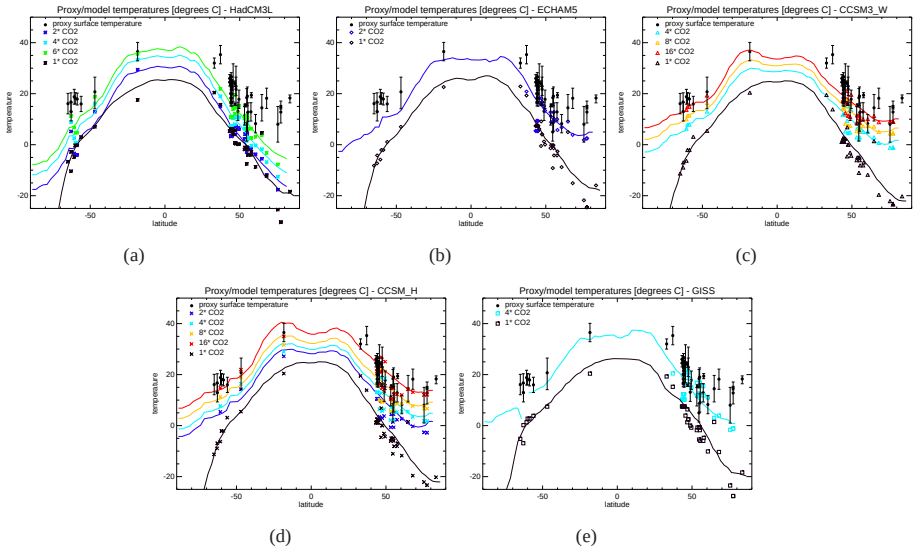


Fig. 7. Comparison of modelled SAT with proxy-derived temperatures, SAT vs. latitude. The simulations at $\times 1$ CO_2 are pre-industrial reference simulations. For the model results, the continous lines represent the zonal mean, and the symbols represent the modelled temperature at the same location (longitude,latitude) as the proxy data. For the proxy data, the symbols represent the proxy temperature, and the error bars represent the range, as given by **?**.

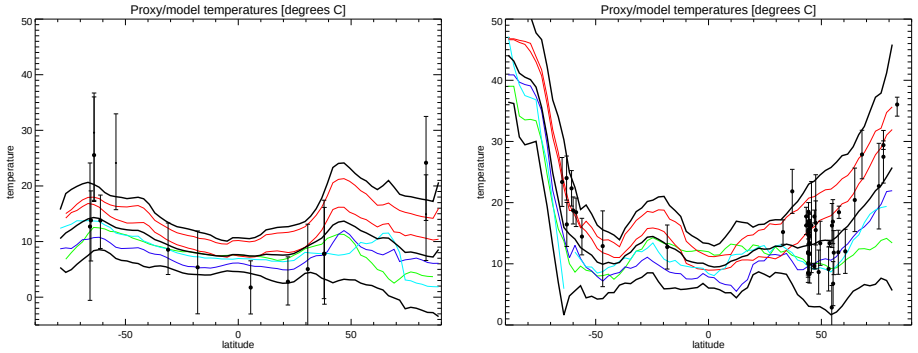
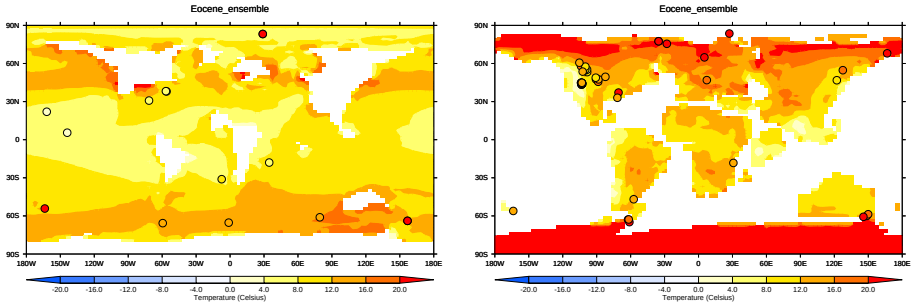
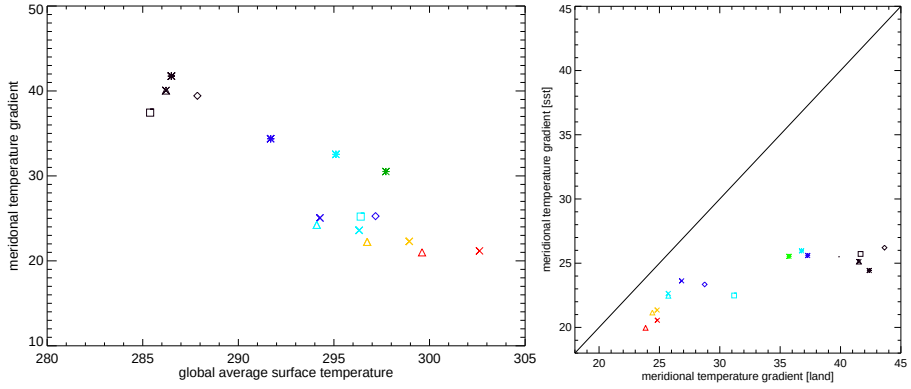


Fig. 8. Zonal ensemble mean model (middle thick black line), and data, presented as an anomaly relative to present/pre-industrial. Outer thick black ~~Light grey model~~ lines indicate $\pm$~~+-~~ 2 standard deviations in the models. Coloured lines represent each individual model simulation in the ensemble, with the colour indicating CO_2 level as in Figures 5 and 7. (a) $[\overline{SST_e} - \overline{SST_p}]$. (b) $[\overline{LAT_e} - \overline{GAT_p}]$. For this Figure, the ~~The~~ ensemble consists of the best simulation from each model, as highlighted in bold in Table 2. Descriptions of the proxy error bars are given in the captions to Figures 5 and 7.



(a) (b)

Fig. 9. Ensemble mean modelled Eocene warming, presented as an anomaly relative to present/pre-industrial. (a) $[SST_e - \overline{SST_p}]$. (b) $[LAT_e - \overline{GAT_p}]$. The ensmble consists of the best simulation from each model, as highlighted in bold in Table 2.



(a) (b)

Fig. 10. (a) Meridional surface temperature gradient $GST_{|\phi|>60} - GST_{|\phi|<30}$, where $|\phi|$ is the absolute value of the latitude in degrees, as a function of global mean surface temperature, $\langle GST \rangle$ for all the simulations presented in this paper. (b) Meridional surface temperature gradient over land / ocean, $SST_{|\phi|>60} - SST_{|\phi|<30}$ *vs.* $LAT_{|\phi|>60} - LAT_{|\phi|<30}$. Symbols and colours correspond to those in Figure 2.

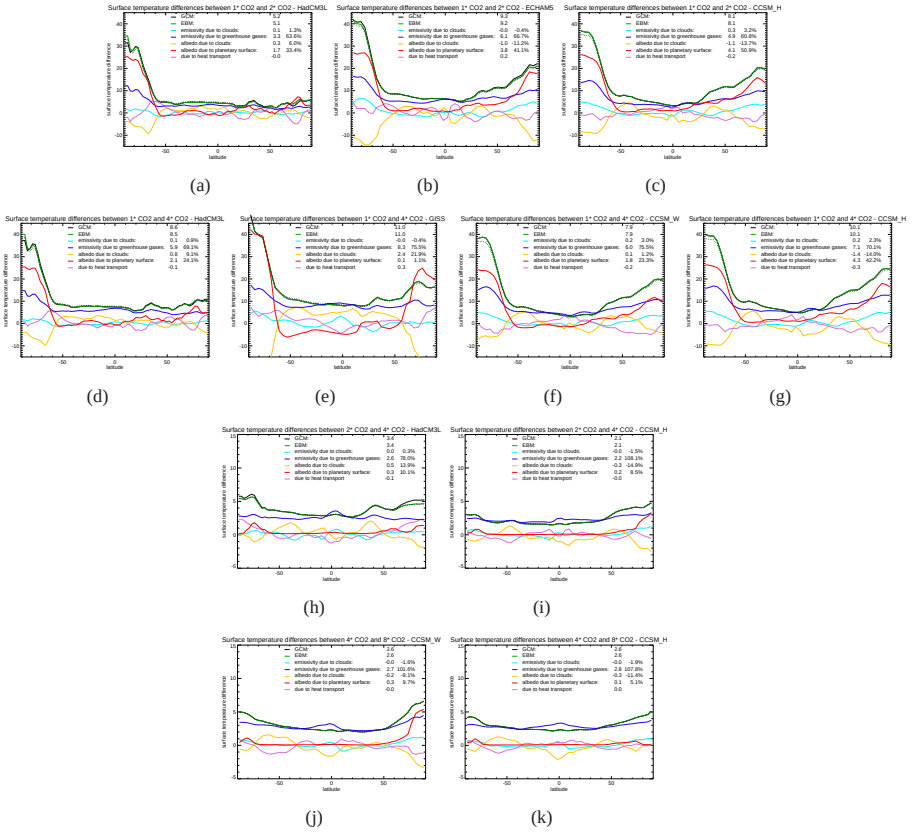


Fig. 11. The zonal-mean surface temperature change under a range of CO_2 transitions, and energy balance analysis of the reasons for the changes. (a-c) $\times 1$ to $\times 2$ CO_2, (d-g) $\times 1$ to $\times 4$ CO_2, (h-i) $\times 2$ to $\times 4$ CO_2, (j-k) $\times 4$ to $\times 8$ CO_2. The simulations at $\times 1$ CO_2 are pre-industrial reference simulations. Note the difference in vertical scale in panels (a-g) compared with (h-k). The dotted lines in the plots show the sum of the various components, which in each case should be very close to the GCM line (i.e. the 'actual' temperature change from the model) and the EBM line (i.e. $\Delta\tau$ as caluclated from Equation 8 for the two climate states).

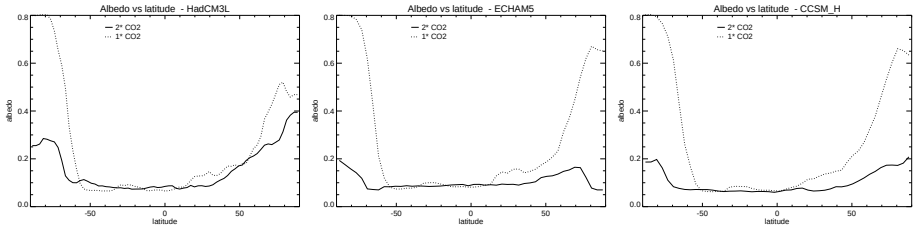


Fig. 12. Zonal mean albedo in the $\times 1$ and $\times 2$ CO_2 simulations using the (a) HadCM, (b) ECHAM, and (c) CCSM_H models. The simulations at $\times 1$ CO_2 are pre-industrial reference simulations.

Table 1. Summary of model simulations in EoMIP. Some models have irregular grids in the atmosphere and/or ocean, or have spectral atmospheres. The atmospheric and ocean resolutions are given in number of gridboxes, X×Y×Z where X is the effective number of gridboxes in the zonal, Y in the meridional, and Z in the vertical. See the original references for more details.

Name	Eocene simulation reference	model name and reference	atmosphere resolution	ocean resolution
HadCM	**?**	HadCM3L, **?**	96×73×19	96×73×20
ECHAM	**?**	ECHAM5/MPI-OM, **?**	96×48×19	142×82×40
CCSM_W	**??**	CCSM3, **??**	96×48×26	100×116×25
CCSM_H	**??**	CCSM3, **??**	96×48×26	100×122×25
GISS	**?**	GISS ModelE-R, **?**	72×45×20	72×45×13

Name	paleogeography	sim. length [years]	CO_2 levels	vegetation	aerosols
HadCM	proprietary ~~propriety~~	>3400	×2,4,6	homogenous shrubland	as control
ECHAM	**?**	2500	×2	homogenous woody savanna	as control
CCSM_W	**?** with marginal sea parameterisation	1500	×4,8,16	**?**	as control
CCSM_H	**?**	>3500 ~~1500~~	×2,4,8,16	**?**	reduced aerosol
GISS	**?**	2000	×2	**?**	as control

Table 2. Global mean temperatures and model mean-error scores for each simulation. Scores are calculated based on the SST (σ_{sst}) and land surface air tempertaure (σ_{lat}) data. Definitions of the scores are given in Equation 2. Rows in bold indicate the best (i.e. lowest σ) CO_2 level for each model.

Model	CO_2	$\langle SST \rangle$	$\langle LAT \rangle$	$\langle GST \rangle$	σ_{sst} [°C]	σ_{lat} [°C]
HadCM	2×	21.45	11.71	18.54	8.8 ~~7.1~~	15.5
	4×	24.19	16.20	21.95	6.0 ~~4.1~~	11.4
	6×	26.25	19.80	24.56	**3.9**	**7.7**
ECHAM	**2×**	24.65	20.59	24.03	**5.8**	**9.7**
CCSM_W	4×	22.24 ~~22.31~~	16.26	20.95	6.7 ~~7.0~~	10.3
	8×	24.45 ~~24.61~~	19.57	23.59	4.0 ~~4.5~~	7.2
	16×	27.14 ~~27.20~~	23.16	26.46	**0.9**	**3.7**
CCSM_H	2×	22.15 ~~22.66~~	15.71	21.12	8.6 ~~7.5~~	11.5
	4×	23.94 ~~24.41~~	18.41	23.17	6.6 ~~5.8~~	8.5
	8×	26.43 ~~26.86~~	21.66	25.79	3.6 ~~3.1~~	5.1
	16×	29.75 ~~30.14~~	26.30	29.47	**0.0**	**0.4**
GISS	**4×**	26.43	21.97	23.25	**3.8**	**6.9**

compared against background early Eocene or pre-PETM SST estimates from the mid- to low latitudes.

We agree that the aggregation of pre-PETM, early Eocene and EECO SST estimates could lead to a misrepresentation of latitudinal temperature gradients and false proxy-proxy mismatches. Given the temporal coverage and age control on the majority of the proxy data, however, it is neither practical nor meaningful to divide all of the data into multiple time-slices. Instead, where there is both reasonable age control and a marked trend in SST estimates between the background early Eocene and EECO, we have divided our estimates along these lines, i.e. into pre-EECO ("background") and EECO ("warm") conditions. Mid- to low-latitude records from the pre-PETM interval are included with the pre-EECO records. As Hollis notes (and see, Hollis et al., 2012), there is evidence for a warming between pre-PETM and post-PETM earliest Eocene conditions in the southern high latitudes but this may be less significant at lower latitudes (Pearson et al., 2007). We do, however, acknowledge that there may be a small cool bias introduced by the use of these records.

Only three marine SST proxy data sets are identified as representing EECO conditions for all or part of the associated SST time series – ODP Site 1172D, Waipara River and the Arctic IODP Site 302-4A. The EECO in ODP Site 1172D record is indentified following Hollis et al. (2012) as spanning ~53.1 to ~49 Ma (588.85 to 562.70 mbsf), with the pre-EECO interval from 54.9 to 53.3 Ma (611.0 to 591.15 mbsf). All of the Waipara River data used here is identified as representing EECO conditions (Hollis et al., 2009; Hollis et al., 2012). Although the early Eocene age model for the Arctic IODP Site 302-4A is poorly constrained between early Eocene hyperthermal events and the termination of the *Azolla* phase, there is a distinct *cooling* in GDGT-derived proxy temperature estimates between the stratigraphically lower (core 27X) and upper (cores 23X to 19X) parts of this section (Sluijs et al., 2008). For the purposes of a more refined proxy-proxy and proxy-model comparison, we have labelled the (warmer) SST data from the lower part of the 302-4A as pre-EECO and the (cooler) SST data from the upper part as EECO. This is in line with speculation in Sluijs et al. (2008) that while there is a global trend of warming through the early Eocene, Arctic SSTs reduced during this interval.

The remaining sites have either relatively poorly constrained age models *and* have no discernable SST trend through the dataset used (ODP Sites 690 and 738), are spot-samples within the early Eocene (Tanzania, Hatchetigbee Bluff) or are well-constrained pre-PETM records. Seymour Island is the exception to this, where there remains uncertainty about even the gross age of this succession. The data used in this compilation is sourced from Telms 3 to 5, which, based on strontium isotope stratigraphy and sparse biostratigraphic data, were thought to span the early Eocene, extending just across the early/middle Eocene boundary (Ivany et al., 2008). A revised age assessment, based on dinoflagellate biostratigraphy, suggests that the lower part of this sequence is middle and not early Eocene in age (Douglas et al., 2011). This new biostratigraphic data remains somewhat tentative, and while awaiting the publication of a fully revised age model for these successions, the Seymour Island data is provisionally included in the SST compilation. It is however, assigned to the background "pre-EECO" category, as it appears to be more likely to be representative of middle Eocene post-EECO cooling.

2) $\delta^{18}O$ records from ODP Sites 690 and 738: Hollis questions the inclusion of planktic foraminferal oxygen isotope records from these two sites based on the well-rehearsed arguments about the fine-scale diagenetic recrystallization of planktic foraminiferal tests and the resetting of oxygen isotope values towards sediment pore-water temperatures (Pearson et al., 2007; Sexton et al., 2006). Part of Hollis's argument is based on the very limited to reversed planktic to benthic $\delta^{18}O$ gradient observed at DSDP Site 277 (Hollis et al., 2009),

which would be indicative of reset planktic foraminiferal $\delta^{18}O$ values. He suggests that the same is true for ODP Sites 690 and 738, a statement which is not entirely accurate. At ODP Site 690 a gradient in $\delta^{18}O$ of ~0.5 to 1.0 ‰ is consistently maintained between the mixed-layer dwelling *Acarinina* and the thermocline-dwelling *Subbotina* throughout the early Eocene (Stott et al., 1990) as well as a mixed layer (*Acarinina*) to benthic (*Cibicidoides*) temperature gradient of ~5ºC (Kennett and Stott, 1990). A similar ~5ºC mixed layer to benthic gradient is also maintained at ODP Site 738 throughout the early Eocene (Barrera and Huber, 1991). Although these planktic to benthic temperature gradients are not large, and probably have been reduced to some degree by planktic foraminifera recrystallization, we argue that the inclusion of these Southern Ocean, high latitude $\delta^{18}O$-derived temperatures is, however, valuable in the context of providing a minimum SST estimate for model-data comparisons. Even taken as minimum temperature estimates, these already constrain most of the model simulations towards the high GHG-forcing end members. We do accept that the warm SST estimates from other localities in the southern high latitudes should not be referred to as "anomalously warm", based solely on a comparison with this planktic foraminifera data.

3) Mg/Ca paleothermometry: In our view, the most uncertain SST estimates included in this compilation are those based on Mg/Ca paleothermometry. This uncertainty stems from 1) the limited understanding of how diagenetic recrystallization of planktic foraminifera effects primary Mg/Ca ratios; and, 2) the absence of a robust estimate of seawater Mg/Ca ratio during the early Paleogene. The low to mid-latitude Mg/Ca temperature estimates included in this compilation are based on the analysis of planktic foraminifera from deep-ocean sites that would be considered substantially recrystalized and inappropriate for oxygen isotope paleothermometry (DSDP Site 527, ODP Sites 865 and 1209). For reasons that are not fully understood, there are indications that foraminiferal Mg/Ca ratios are less susceptible to diagenetic alteration than $\delta^{18}O$, and that the potential bias is towards higher temperatures by the replacement of primary biogenic calcite with higher Mg diagenetic calcite (Sexton et al., 2006). While awaiting a fuller understanding of the impacts of recrystallization on Mg/Ca temperature estimation, we tentatively opt to include the available Mg/Ca data, even if from poorly preserved foraminifera tests, as a more robust assessment of this data may be possible in future.

Hollis suggests the use of an early Paleogene seawater Mg/Ca ratio of 4 or a range of 4 to 5 mol mol-1, arguing that a higher value produces a closer match between oxygen isotope and Mg/Ca paleothermometry from Eocene planktic foraminfera (Sexton et al., 2006) and a better match between the TEX_{86} and Mg/Ca SST estimates from the Southwest Pacific used in this study. Both of these observations rely on proxy-proxy comparisons and risk introducing false agreement between proxies, which have, in effect been calibrated against each other in ancient material. Notwithstanding this, inferences from the long-term benthic foraminfera $\delta^{18}O$ and Mg/Ca records are that early Paleogene seawater Mg/Ca is <4 mol mol^{-1} (Cramer et al., 2011; Lear et al., 2002). These lower values are more in line with both independent modelling studies (Farkaš et al., 2007) and estimates from the geochemistry of ridge flank calcium carbonate veins (Coggon et al., 2010). The range 3 to 4 mol mol^{-1} also incorporates the values used by recent early Paleogene Mg/Ca paleothermometry studies – 3.19 mol mol^{-1} (Tripati and Elderfield, 2004), 3.35 mol mol^{-1} (Creech et al., 2010; Hollis et al., 2009) and 4 mol mol^{-1} (Hollis et al., 2012). Although we eagerly await a better constraint on seawater Mg/Ca, and note that future studies may place this value outside of the range we use here, 3-5 mol mol^{-1} remains our favoured best estimate for the likely range of early Paleogene seawater Mg/Ca.

On the more specific point questioning the estimate of early Eocene seawater Mg/Ca of 3.5 mol mol^{-1} based on Lear et al. (2002), the text requires clarification. It should state that this

value is calculated for the purposes of this study *using* the Lear et al. (2002) preferred calibration for *O. umbonatus* (A=0.114; B=1.008), a foraminiferal Mg/Ca ratio of 2.78 mmol mol^{-1} and seawater temperature estimate of 12.4ºC at 49 Ma.

4) Seawater $\delta^{18}O$ estimation: Following the suggestion of Hollis, we have added the Roberts et al (2011) modelled $d^{18}O_{sw}$ in addition to our current values from Zachos et al and Tindall et al. Unsurprisingly, this increases the range of temperatures reconstructed. The (new) Figure 1 shows clearly, however, that the uncertainties introduced from the different sweater estimates are not the dominant uncertainty at any of the sites. N.B. Chris Roberts is now a co-author on this paper.

5) $\delta^{18}O$–temperature calibration: On consideration, and following the comments of Hollis, we have recalculated $\delta^{18}O$-based SSTs using the calibrations for the symbiotic planktic foraminifera, *Orbulina universa*, under both high and low light conditions (equations 1 and 2) of Bemis et al. (1998) rather than the calibration of Erez and Luz (1983). Together, these two equations bracket most of the variability in planktic foraminifera temperature-$\delta^{18}O$ space within modern plankton tow data (Bemis et al., 1998). The standard error on equation 1 (low light) and 2 (high light) are ±0.7 and 0.5ºC respectively.

6) TEX_{86} paleotemperature estimates: We acknowledge the concerns of Hollis over the use of all three GDGT-based proxies to produce a single range and median SST value for each site. As Hollis notes, the use of both TEX_{86}^{H} and $1/TEX_{86}$ at high latitudes, which give similar SST estimates that are normally somewhat warmer than TEX_{86}^{L}, risks introducing a warm-bias to the derived median SST value. Likewise, we accept that there is significant uncertainty about the application of TEX_{86}^{L} at low and mid-latitudes. Although this application certainly lies outside of the intentions of its original proponents, as with its recent use as proxy of choice for records from the Southwest Pacific (Hollis et al., 2012), it may in future also prove suitable for SST estimation in other specific paleoenvironments. To address these concerns, and make the compiled SST dataset both transparent in its current presentation and more robust to developments in GDGT paleothermometry, we have now plotted each of the three GDGT-based proxies independently at all sites. This allows a site-by-site model-data comparison that is responsive to the changing understanding of GDGT-based proxies. The only location where this is not possible is Tanzania, where TEX_{86}^{L} produces clearly erroneous results on samples with BIT indices between 0.3 and 0.5, and is excluded. As Hollis notes, the errors on the GDGT based proxies should be ±2.5 for TEX_{86}^{H} (GDGT index-2), ±4.0 for TEX_{86}^{L} (GDGT index-1) and ±5.4 for 1/TEX (Kim et al., 2010).

7) terrestrial proxies: The reviewer is correct that we are using data direct from this recent peer-reviewed dataset, and have not modified it. Our marine dataset is discussed in more detail, but this is precisely because it has been compiled specifically for this paper. We do not feel that there is scope in this paper to repeat the analysis and discussion from Huber and Caballero. It is worth noting that there are fundamentally different approaches to creating a compilation for the purpose of comparing with climate model output for a given time slice and the approach one might use to create a high quality, trend-oriented time series for studying evolution of climate through a time interval. With infinite and perfect time series these approaches converge, but for patchy sampling with uncertain data the strategies differ.

In the model-data comparison case, in which one is trying to characterize an entire model grid cell, a region 200kmx200km for a time slice that is 5-10 million years long, then one wants to keep as much of the spatial heterogeneity as possible. Each model grid cell has a constant elevation and constant climate, whereas each locality likely represents a tiny microcosm of

the larger terrain and it is necessary to take a number of these averaged together to approach something at all like the mean climate of the model grid cell. So, it is correct to average together the rainforest and the savannah if they are near each other, because that is the average climate of the grid cell. You also want to average together floras that are 5-10 million years apart since we need all the information in the temporal domain we can to get anything close to a robust time mean.

In addition, Matt Huber has compiled "early", "middle" and "late" Eocene terrestrial data separately and no matter how these are defined, and no matter how they are restricted (i.e. including hyperthermals or the MECO or excluding them), there is no statistically significant difference in MAT between them.

8) Additional minor points:

1. [2.3] CCSM3_H was developed for Hollis et al. (2009) not for Liu et al. (2009), which has a late Eocene model.

The references for the CCSM_H runs are correct. Hollis et al (2009) had the 2240 case but not run out to its full length.

2. [3.1.1] Bemis et al. (1998) showed that the Erez and Luz (1983) equation suffers from a warm bias of up to 3.5°C and introduced alternative equations, which give values very similar to Kim and O'Neil (1997). Why were these equations not used at least as an alternative to Erez and Luz. I cannot find reference to a standard error of 1.43°C in the latter paper, on p. 1028 they list sources of error that total 2.15°C.

Done – we now use the Bemis calibration. [See (5), above].

3. [3.1.3] Line 18 - "high" and "low" should be reversed. Line 27 – delete (TEX86 and TEX86L)?

Done.

Line 3(p. 1239) – Hollis et al., 2012 not in refs (replace with Hollis et al., 2011 – see below).

The Hollis reference has now been updated.

Line 11- The error on each of three proxies is different, 2.5 for TEX86H, 4 for TEX86L and 5 for 1/TEX86. Why is the minimum error of 2.5 used here?

Done. [See (6), above].

4. [3.2] Give simple explanation of what LMA and CLAMP are – physiognomic analysis of leaf fossils … Note the second paragraph is out of place, should go at the end of [3] or the start of [4].

Done.

5. [Figure 1] Model labels overprinted

We leave the figure as-is. To move the labels around would be confusing as each one is currently next to the corresponding data point.

6. [Figure 4] "temporal uncertainty (black bar) and calibration uncertainty (grey bar)"

Done. Error bars now combine temporal and calibration uncertainty, and Figure 1 shows the various assumptions about ocean water composition.

7. Global find and replace "New Zealand" with "southwest Pacific"!

Done!

9) Presentation Quality:

The figures are all appropriate although in several cases they suffer from being too small to be legible on a standard screen and are certainly too small for A4 printing. It is especially hard to resolve the colours for proxy data in Figures 2 and 3.

We feel that these figures are appropriate. They are pdfs so any online reader can zoom in on the plots ad infinitum.

The marine proxy table lacks references and would benefit from comments on individual records where the quality of data (ODP sites 690, 738; Hatchetigbee TEX86) or the age of the record is questionable (Seymour Island).

Done.

The terrestrial proxy table is superfluous as it is simply an unlabelled extract from the table in Huber and Caballero (2011)

We disagree – Huber and Caballero is just a pdf, making extraction of actual values more problematic than using our spreadsheet.

Barrera, E. and Huber, B.T., 1991. Paleogene and early Neogene oceanography of the southern Indian Ocean: Leg 119 foraminifera stable isotope results. Proceedings of the Ocean Drilling Program, Scientific Results, 119: 693-717.

Bemis, B., Spero, H., Bijma, J. and Lea, D., 1998. Reevaluation of the oxygen isotopic composition of planktonic foraminifera: Experimental results and revised paleotemperature equations. Paleoceanography, 13(2): 150-160.

Bijl, P.K. et al., 2009. Early Palaeogene temperature evolution of the southwest Pacific Ocean. Nature, 461: 776-779.

Coggon, R.M., Teagle, D.A.H., Smith-Duque, C.E., Alt, J.C. and Cooper, M.J., 2010. Reconstructing past seawater Mg/Ca and Sr/Ca from mid-ocean ridge flank calcium carbonate veins. Science, 327: 1114-1117.

Cramer, B.S., Miller, K.G., Barrett, P.J. and Wright, J.D., 2011. Late Cretaceous-Neogene trends in deep ocean temperature and continental ice volume: Reconciling records of benthic foraminiferal geochemistry ($\delta^{18}O$ and Mg/Ca) with sea level history. Journal of Geophysical Research, 116(C12023): doi:10.1029/2011JC007255.

Creech, J.B., Baker, J.A., Hollis, C.J., Morgans, H.E.G. and Smith, E.G.C., 2010. Eocene sea temperatures for the mid-latitude southwest Pacific from Mg/Ca ratios in planktonic and benthic foraminifera. Earth and planetary science letters, 299(3-4): 483-495.

Douglas, P.M. et al., 2011. Eocene Southern high latitude sea surface temperatures: new constraints from clumped isotope paleothermometry, AGU fall meeting, San Francisco, pp. PP33B-1924.

Erez, J. and Luz, B., 1983. Experimental paleotemperature equation for planktonic foraminifera. Geochemica et Cosmochemica Acta, 47: 1025-1031.

Farkaš, J. et al., 2007. Calcium isotope record of Phanerozoic oceans: Implications for chemical evolution of seawater and its causative mechanisms. Geochemica et Cosmochemica Acta, 71: 5117-5134.

Hollis, C.J. et al., 2009. Tropical sea temperatures in the high-latitude South Pacific during the Eocene. Geology, 37(2): 99-102.

Hollis, C.J. et al., 2012. Early Paleogene temperature history of the Southwest Pacific Ocean: reconcilling proxies and models. EPSL, 349-350: 53-66.

Ivany, L.C. et al., 2008. Eocene climate record of a high southern latitude continental shelf: Seymour Island, Antarctica. Geological Society of America Bulletin, 120(5/6): 659-678.

Kennett, J.P. and Stott, L.D., 1990. Proteus and Proto-Oceanus: ancestral Paleogene oceans as revealed from antarctic stable isotopic results: ODP Leg 113. Proceedings of the Ocean Drilling Program, Scientific Results, 113: 865-880.

Kim, J.-H. et al., 2010. New indices and calibrations derived from the distribution of crenarchaeal isoprenoid tetraether lipids: Implications for past sea surface temperature reconstructions. Geochimica et Cosmochimica Acta, 74(16): 4639-4654.

Lear, C., Rosenthal, Y. and Slowey, N., 2002. Benthic foraminiferal Mg/Ca-paleothermometry: A revised core-top calibration. Geochimica et Cosmochimica Acta, 66(19): 3375-3387.

Pearson, P.N. et al., 2007. Stable warm tropical climate through the Eocene Epoch. Geology, 35(3): 211-214.

Roberts, C.D., LeGrande, A.N. and Tripati, A.K., 2009. Climate sensitivity to Arctic seaway restriction during the early Paleogene. Earth and Planetary Science Letters, 286: 576-585.

Roberts, C.D., LeGrande, A.N. and Tripati, A.K., 2011. Sensitivity of seawater oxygen isotopes to climatic and tectonic boundary conditions in an early Paleogene simulation with GISS ModelE-R. Paleoceanography, 26(PA4203): doi:10.1029/2010PA002025.

Sexton, P., Wilson, P. and Pearson, P., 2006. Microstructural and geochemical perspectives on planktic foraminiferal preservation:"Glassy" versus "Frosty". Geochemistry Geophysics Geosystems, 7(12): Q12P19.

Sluijs, A. et al., 2008. Arctic late Paleocene-early Eocene paleoenvironments with special emphasis on the Paleocene-Eocene thermal maximum (Lomonosov Ridge, Integrated Ocean Drilling Program Expedition 302). Paleoceanography, 23(1).

Stott, L.D., Kennett, J.P., Shackleton, N.J. and Corfield, R.M., 1990. 48. The evolution of Antarctic surface waters during the Paleogene: inferences from the stable isotopic composition of planktonic foraminifers, ODP Leg 113. Proceedings of the Ocean Drilling Program, Scientific Results, 113: 849-863.

Tindall, J. et al., 2010. Modelling the oxygen isotope distribution of ancient seawater using a coupled ocean–atmosphere GCM: Implications for reconstructing early Eocene climate. Earth and Planetary Science Letters, 292(3-4): 265-273.

Tripati, A. and Elderfield, H., 2004. Abrupt hydrographic changes in the equatorial Pacific and subtropical Atlantic from foraminiferal Mg/Ca indicate greenhouse origin for the thermal maximum at the Paleocene-Eocene Boundary. Geochemistry Geophysics Geosystems, 5(2): Q02006.

Zachos, J.C., Dickens, G.R. and Zeebe, R.E., 2008. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. Nature, 451: 279-283.

Zachos, J.C., Stott, L.D. and Lohmann, K.C., 1994. Evolution of early Cenozoic marine temperatures. Paleoceanography, 9(2): 353-387.