The interpretation and use of biases in decadal climate predictions

Ed Hawkins*, Buwen Dong, Jon Robson, and Rowan Sutton

NCAS-Climate, Department of Meteorology, University of Reading, UK

Doug Smith

Met Office Hadley Centre, Exeter, UK

Accepted by Journal of Climate
March 19, 2014

* *Corresponding author address:* Ed Hawkins, Department of Meteorology, University of Reading, Reading. RG6 6BB. UK. E-mail: e.hawkins@reading.ac.uk

ABSTRACT

Decadal climate predictions exhibit large biases, which are often subtracted and forgotten. However, understanding the causes of bias is essential to guide efforts to improve prediction systems, and may offer additional benefits. Here we investigate the origins of biases in decadal predictions, and whether analysis of these biases might provide useful information. We focus especially on the lead time dependent bias tendency. We initially develop a 'toy' model of a prediction system and use it to show that there are several distinct contributions to bias tendency. Contributions from sampling of internal variability and a start-time dependent forcing bias can be estimated and removed to obtain a much improved estimate of the true bias tendency, which can provide information about errors in the underlying model and/or errors in the specification of forcings. We argue that it is the true bias tendency, not the total bias tendency, that should be used to adjust decadal forecasts.

We apply the methods developed to decadal hindcasts of global mean temperature made using the HadCM3 climate model, and find that this model exhibits a small positive bias tendency in the ensemble mean. When considering different model versions we show that the true bias tendency is very highly correlated with both the Transient Climate Response (TCR) and non-greenhouse gas forcing trends, and can therefore be used to obtain observationally constrained estimates of these relevant physical quantities.

1. Introduction

Until recently, projections of future climate have been generated by running climate models forced by estimates of future natural and anthropogenic (e.g. from greenhouse gases and aerosols) radiative forcing. The motivation for decadal climate predictions is to improve on these standard projections by using observations to initialise predictable modes of natural variability, and by correcting errors in a model's response to past radiative forcings. Producing climate predictions that are initialised using observations of the current climate state is now a major field of scientific research (e.g., Smith et al. 2007; Keenlyside et al. 2008; Pohlmann et al. 2009; Smith et al. 2013). For example, initialised decadal climate prediction experiments are a major component of phase 5 of the Coupled Model Intercomparison Project (CMIP5; Meehl et al. 2009; Taylor et al. 2012; Meehl et al. 2013). Decadal climate predictions could potentially be of great benefit to society, for example helping to inform decisions on adaptation to a changing climate. However, there are many challenges in producing forecasts that are useful for adaptation decisions (e.g. Meehl et al. 2009; Oreskes et al. 2010).

One key challenge in producing robust predictions of future climate is to demonstrate an ability to make predictions in the past ('hindcasts'). Comparisons between hindcasts and past observations offer a wealth of information for assessing the strengths and weaknesses of a prediction system, including information that can guide work to improve the system. Such an approach has proved invaluable in weather forecasting (e.g. Ferranti and Viterbo 2006). Comparisons may focus on specific case studies (e.g. Robson et al. 2012, Yeager et al. 2012), particular regions (e.g. Toniazzo and Woolnough 2013) or on the average behaviour of a system over a longer period (e.g. Smith et al. 2007, 2010; van Oldenborgh et al. 2012). A particularly important issue for decadal climate predictions is the existence of large *biases*, i.e. systematic differences between hindcasts and observations. Biases may vary with the lead time of hindcasts and are often larger than the anomalies that the system is aiming to predict. In this situation the current standard approach (e.g. Goddard et al. 2013) is to subtract the mean bias from all hindcasts before assessing other aspects of the system performance (e.g. RMSE). Such an approach is pragmatic but assumes a linear additivity between bias and forced response and ignores many important issues, such as: Why is the bias present? Does it provide any useful information? Could it be reduced?

The aim of this paper is to investigate the first two of these questions in particular, initially in the context of an idealised "toy" model, and secondly using results from a real decadal prediction system. We focus especially on the growth of bias with lead time, which we demonstrate offers valuable information about a prediction system and the underlying climate model. We then show further that analysis of biases for different model versions can be used to obtain useful information about the real world, in particular new constraints on the Transient Climate Response (TCR), which measures the transient sensitivity of the climate system to increases in greenhouse gases.

The structure of the paper is as follows. Section 2 discusses the design of decadal prediction experiments, and clarifies terminology. Section 3 introduces our toy model of a decadal prediction system, explains how the bias can be decomposed into distinct contributions, and examines sampling issues. The methodology we develop is then applied to predictions of global mean surface air temperature from an operational decadal prediction system in Sections 4 and 5. Conclusions and discussion of implications are in Section 6.

2. Experimental design and terminology

There are several types of decadal climate prediction experiment discussed in the literature. One important issue is the specification of external radiative forcings in the hindcasts. The two main choices are:

- 'Projection'-type - Anthropogenic forcings are assumed to be known, but 'projected' natural forcings are used (e.g. see Smith et al. 2007). In this case any volcanic aerosol present at the forecast start time is allowed to decay, but no 'future' volcanic aerosol is used. In addition, the solar cycle is repeated from the previous cycle. This approach attempts to mimic the realistic situation in which there is little knowledge of future natural forcing.

- ’CMIP5’-type - All forcings are assumed to be known. This is the design adopted by the CMIP5 protocol (Taylor et al. 2012).

In addition, hindcasts may be initialised using observations at the forecast start time (’Assim’ - because assimilation is used to generate the initial states), or be initialised directly from a model state without the use of observations (’NoAssim’).

The simplest case is arguably the ‘NoAssim CMIP5’ type, corresponding to traditional so-called “transient” climate model simulations. However, the ensemble sizes for these simulations tend to be small (fewer than 5), which - as we will show - limits the robustness of the bias analysis. In this study we focus on the ’NoAssim projection’ type of hindcasts, as performed by the UK Met Office (see Smith et al. 2007, hereafter S07). The Met Office used this approach to produce a very large ensemble of hindcasts with different versions of the same GCM (Smith et al. 2010), which proves to be a very useful resource for our analysis. However, in examining these hindcasts we must take into account the difference between the natural forcings used to force the model and those that occurred in the real world.

The reason that we focus on ’NoAssim’-type experiments is that understanding the biases in these experiments is a pre-requisite for understanding the biases in ’Assim’-type experiments. We demonstrate that the bias derived from ’NoAssim’ experiments provides useful information, and we will be investigating applications to ’Assim’-type experiments in future work.

3. Estimating bias in a toy model of a decadal prediction system

We first build a toy model of a decadal prediction system to examine some of the issues involved with estimating the bias of a real prediction system.

a. Bias of hindcasts

Pseudo-observations, $O(t)$, are generated by assuming an externally forced linear trend in time, with added red noise,

$$O(t) = \widetilde{O} + \alpha t + \epsilon(t), \tag{1}$$

where t is time, $\widetilde{O}$ is the ‘observed’ climatology, α is the slope of the linear trend, and the red noise is denoted by $\epsilon(t)$.

We first assume that the ensemble mean of our pseudo-hindcasts (N) for the same quantity can be generally represented, for start time T and lead time τ, by

$$N(T, \tau) = \widetilde{N} + (T + \tau)\gamma \tag{2}$$

where $\widetilde{N}$ is the model climatology and γ is the modelled linear response to the external forcing. If $\alpha \neq \gamma$ then the climate model would produce a different trend from the observations and therefore be biased. This could either be because the model is in error, or because there is an error in the specification of the forcing (see later). This equation for N assumes that we have an infinite ensemble of hindcasts, as there is no noise in the ensemble mean. This assumption will be relaxed later. Note that these pseudo-hindcasts are only attempting to predict the forced response, and not the internal variability component.

The bias (B) of a prediction system is simply the mean error as a function of prediction lead time,

$$B(\tau) = \frac{1}{L}\sum_{T=1}^{L}\left[N(T, \tau) - O(T + \tau)\right] \tag{3}$$

where L is the number of hindcast start dates and we assume that there is a decadal hindcast ($\tau = 1$ to 10 years) started every year between, and including, $T = 1$ and $T = L$. Note that in an operational system N and O would often represent anomalies from a particular reference period. However, our analysis focusses on 'bias tendency' (defined below) which is independent of the choice of reference period.

b. Correcting the bias for observed variability

The estimated bias defined in Eqn. 3 has two contributing factors, namely the true bias (if $\alpha \neq \gamma$ or $\widetilde{N} \neq \widetilde{O}$) and a bias from an insufficient sampling of the internal variability in the observations. Ideally, we would like to correct for this second variability contribution to obtain the true bias.

Following Robson (2010), in the case of an infinite ensemble in a stationary climate ($\alpha = \gamma = 0$), the bias from Eqn. 3 would be,

$$B_{\text{stationary}}(\tau) = \frac{1}{L}\sum_{T=1}^{L}\left[\widetilde{N} - \widetilde{O} - \epsilon(t)\right], \tag{4}$$

$$= \widetilde{N} - \widetilde{O} - \frac{1}{L}\sum_{t=\tau}^{L+\tau}\epsilon(t), \tag{5}$$

$$= \widetilde{N} - \widetilde{O} + B_{\text{obsvar}}(\tau), \tag{6}$$

where t represents time and $B_{\text{obsvar}}(\tau)$ is the mean of the observational anomalies used for validation for a particular lead time τ. An important point is that different observations are used for different lead times. Thus $B_{\text{obsvar}}(\tau)$ is an estimate of the bias due to the insufficient sampling of the observed variability and will tend to zero as L increases leaving the true bias, $\widetilde{N} - \widetilde{O}$.

For the more realistic case when the climate is not stationary, and there is a trend in the observations ($\alpha \neq 0$) then we can estimate:

$$B_{\text{obsvar}}(\tau) = -\frac{1}{L}\sum_{t=\tau}^{L+\tau} \text{detrended}[O(t)], \tag{7}$$

and this is the definition we adopt. In the toy model examples shown here we use a linear detrending. When considering the real observations we performed sensitivity tests to explore linear and quadratic detrending and the results were very similar (not shown), so assume a linear detrending in all that follows.

A schematic demonstrating B_{obsvar} for different lead times is shown in Fig. 1 with pseudo-observations in black, which include a linear trend and red noise, and some predictions (for a non-infinite ensemble) shown in red in each panel. The grey regions indicate the area to be integrated to give the value of B_{obsvar}, which varies with the lead time chosen, and need not be zero, as shown in Fig. 1d.

c. *Bias tendency*

In this analysis we generally consider the 'bias tendency' (B') rather than the bias itself, i.e. we use the bias relative to the bias for the mean of the first year,

$$B'(\tau) = B(\tau) - B(\tau = 1). \tag{8}$$

This choice is made because we want to consider the growth of bias with lead time, which is natural for a prediction system. We do not use $\tau = 0$ to avoid arbitrary assumptions about defining climatological periods. Hence, this bias tendency has the desirable property of being independent of the choice of climatology.

Similarly to the bias, the observed variability correction is also made into a tendency,

$$B'_{\text{obsvar}}(\tau) = B_{\text{obsvar}}(\tau) - B_{\text{obsvar}}(\tau = 1), \tag{9}$$

as shown in Fig. 1e, and an estimate of the underlying *true bias tendency* (B'_{true}) is then,

$$B'_{\text{true}}(\tau) = B'(\tau) - B'_{\text{obsvar}}(\tau). \tag{10}$$

The nature of the bias growth may give valuable information about the physical processes which cause prediction error, potentially allowing particular parameterisations to be targeted for improvement, for example.

d. Estimating the bias tendency in the toy hindcasts

To test the bias tendency estimates described above, we first consider whether we can estimate the true bias tendency of the toy model using various numbers of hindcast start dates. Here, we generally assume that $\alpha = 0.016$K/year and that the red noise in Eqn. 1 (ϵ) has an AR(1) parameter, $\beta = 0.5$, and total variance, $\sigma_\epsilon^2 = 0.01$K. These values are chosen to roughly simulate observed annual global mean surface air temperature (SAT) observations since 1850 (Brohan et al. 2006), although the conclusions are insensitive to the exact choices. We pick $\gamma = 0.020$K/year, i.e. the toy hindcasts are positively biased by 25%, and retain the infinite ensemble assumption for now.

An example of such a hindcast system is shown in Fig. 2a for decadal hindcasts started every year for $L = 20$ years, where the black line represents the observations, the solid blue line is the true forced trend (α), the dashed blue line is a linear fit to the observations used in the estimation of B'_{obsvar}, and the red lines represent the pseudo-hindcasts (N) which are identical because of the infinite ensemble assumption.

In Fig. 2b, we show estimates of the bias tendency for the situation in Fig. 2a. The solid blue line uses the definition of uncorrected bias tendency (Eqn. 8), and the dashed blue line corrects for the observed variability using Eqn. 10. Note that the dashed blue line does not match the true bias (grey shading) because the estimated trend from the observations is not correct, i.e. the estimate of B'_{obsvar} is not exact. If the true forced trend is used in the estimation of B'_{obsvar} then the true bias tendency is recovered (black line).

We next simulate 1000 realisations of the pseudo-observations and hindcast sets. Bias tendency estimates for 10 examples of these realisations are shown in Fig. 2c. With these 20 start dates there is a wide range of estimated bias tendencies. For different numbers of hindcast start dates (L), Fig. 3 demonstrates that correcting the bias tendency using B'_{obsvar} (dashed line) reduces the error in the estimates of bias tendency at a lead time of 10 years compared to using the uncorrected bias tendency (solid line). Both estimators of the bias tendency are themselves unbiased, i.e. the mean over all realisations equals the true bias tendency (not shown). The spread in bias tendency estimates decreases with the number of start dates as more observations allow more accurate estimates. The observed variability correction also becomes smaller with more start dates. When analysing the operational NoAssim hindcasts in Section 4 we generally use 40 start dates, so the spread is around half as large as suggested in Fig. 2c.

For the particular set of toy model parameters chosen here, we see that the expected error in the bias tendency estimate becomes smaller than the bias itself (grey line in Fig. 3), i.e. the sign of the true bias tendency could be detected, for around $L = 15 - 20$ hindcast start dates. For fewer hindcasts, the uncertainty in the bias estimates does not allow a detection, with the implication for ensemble design that more start dates are required. If the bias is uncorrected then more start dates are required to detect the bias.

e. Forcing bias and consistent verification times

So far we have assumed that the radiative forcing that is causing a warming or cooling trend has been correctly specified and so any bias tendency is due to errors in the model response to this forcing. However, there are two types of forcing bias which could make this assumption invalid, namely start-time independent and start-time dependent bias. The 'CMIP5' design discussed in Section 2 results in start-time independent forcing biases because all hindcasts see the same forcing at the same date. However, for the 'Projection' design this is not the case: hindcasts started from different dates may see different forcings. For example, a hindcast started in 1989 would not include any volcanic aerosol from the Mount Pinatubo eruption in 1991, whereas a hindcast started in 1992 would. Thus there is a start-time dependent forcing bias. S07 noted that this type of forcing bias makes a significant contribution to the bias of a set of hindcasts. They attempted to remove it, somewhat arbitrarily, by excluding years just after volcanic eruptions from the estimation of the bias. Fortunately, a further correction is available to account for this start-time dependent bias.

In deriving, B from Eqn. 3 we chose to use all possible combinations of start dates and verification times. However, an alternative is to use a 'consistent' set of verification times, which only includes years where all lead times, τ, can be *simultaneously* assessed, i.e. the *same* observation can be used to assess the bias at all lead times. In the schematic of Fig. 1 these times are shown by the range of the blue bars, i.e. years 11-21 in this example, as year 11 is the earliest time that a 10 year lead time forecast can be verified (along with forecasts for lead times of 1-9 years), and year 21 is the last time that a 1 year lead time can be verified (along with forecasts for lead times of 2-10 years).

Using these consistent verification times, assuming there is no start time dependent forcing bias and an infinite ensemble, and generalizing from Eqn. 3, the bias becomes,

$$
\begin{aligned}
B_{\text{consis}}(\tau) &= \frac{1}{L-\tau_{\max}+1}\sum_{t=1+\tau_{\max}}^{L+1}\left[N(T,\tau)-O(t)\right], &(11)\\
&= \frac{1}{L-\tau_{\max}+1}\sum_{t=1+\tau_{\max}}^{L+1}\left[N(t)-O(t)\right], &(12)\\
&= A, &(13)
\end{aligned}
$$

where $\tau_{\max}$ is the largest lead time to be considered. Crucially, for this particular choice of verification times, all the terms on the right hand side of Eqn. 12 are *independent* of lead time, because $N(t)$ is the same for all lead times and B'_{obsvar} is zero for this choice of verification times (Fig. 1). In this instance, $B_{\text{consis}}(\tau)$ is a constant (A) with lead time, and therefore, the bias tendency using consistent verification times is,

$$B'_{\text{consis}}(\tau) = B_{\text{consis}}(\tau) - B_{\text{consis}}(1), \tag{14}$$

$$= 0. \tag{15}$$

Hence, *in the absence of a start-time dependent forcing bias*, B'_{consis} is exactly zero (assuming an infinite ensemble).

To test the impact of a start-time dependent forcing bias in our toy model, we generalise Eqn. 1 by adding a volcanic eruption into the pseudo-observations, within the consistent validation time period, of the form,

$$V(\xi) = 0.2 \exp(-\xi) \tag{16}$$

where V is the temperature response to a volcanic eruption, which reduces over time (ξ, measured in years) with an exponential decay timescale of 1 year, from a peak impact of 0.2K. We also assume that the hindcasts also include this impact, but only after the eruption has occurred.

Repeating our toy hindcasts (Fig. 4), still assuming an infinite ensemble, demonstrates that the measured bias tendency (blue) is over-estimated when compared to the true bias tendency (dark grey), because the bias tendency due to the volcanic eruption is non-zero (light grey). B'_{consis} is shown by the red line in Fig. 4b, which matches the forcing bias tendency (light grey) as expected.

Note especially that to estimate B'_{consis} from the data there is no need to assume any functional form for the forcing bias. Therefore, we can correct for the start-time dependent forcing bias by estimating the bias tendency using all verification times, and subtracting off the bias tendency estimated using consistent verification times (B'_{consis}). Generalising Eqn. 10,

$$B'_{\text{true}}(\tau) = B'(\tau) - B'_{\text{obsvar}}(\tau) - B'_{\text{consis}}(\tau). \tag{17}$$

The green lines in Fig. 4b are an example of such an estimate using the bias tendency corrected only by the consistent verification times (solid) and using Eqn. 17 (dashed). Below we will demonstrate that it is necessary to remove the forcing bias in this way to obtain a robust estimate of the true bias tendency, which is the key quantity of interest.

We note here that there are still two contributions to the true bias tendency. The first is errors in the underlying climate model; for example, if the sensitivity of the model to greenhouse gas forcing is higher or lower than that of the real world, the hindcasts will warm too rapidly or too slowly, giving a positive or negative bias tendency. The second is (start-time independent) errors in the forcing applied to the model; for example, if the negative radiative forcing due to anthropogenic aerosols is lower or higher in the model than in the real world, this will also give a positive or negative bias tendency. Correcting the bias tendency using the period of consistent verification times does not deal with the issue of forcing errors that may occur outside of the period of consistent verification times, and this is discussed further when considering the real observations.

Finally, it should be noted that estimating the bias tendency using all verification times and subtracting off the bias tendency using consistent verification times is not the same as estimating the bias tendency using 'non-consistent' verification times (not shown).

f. How many ensemble members are needed?

As discussed above, we have so far assumed that the toy hindcasts have infinite ensemble members. We now relax this assumption to understand how many ensemble members would be required to ensure a robust bias tendency estimate.

For a finite ensemble, our toy model for the predictions is generalized from Eqn. 2 to,

$$N(T,\tau) = (T+\tau)\gamma + \zeta(T,\tau) \tag{18}$$

where ζ is red noise with the same AR1 parameter as the pseudo-observations ($\beta = 0.5$) and a noise component which depends on M, the number of ensemble members, i.e. $\sigma(\zeta) = \sigma_\epsilon/\sqrt{M}$. Note that this definition is equivalent to taking the mean of M different ensemble members, each with variance σ_ϵ^2.

Fig. 5 explores the spread in estimates of the true bias tendency using various values for M, making (or not) the different corrections discussed above. This spread is derived from 100,000 different realisations of the toy model. The colours represent using 20 start dates (grey) and 40 start dates (blue). Firstly, the most reliable and accurate estimate of the true bias is when all the corrections described above are applied (Fig. 5a). For the other cases, the bias estimate itself becomes more biased, or more uncertain (Fig. 5b,c,d).

In addition, as the number of ensemble members is increased the uncertainty in the bias estimates initially decrease, but then stabilise. For $M \gtrsim 8$, the expected error in the bias remains roughly constant. This analysis suggests that as long as $M \gtrsim 8$, then the ensemble is effectively infinite for global mean temperature. In addition, to detect the sign of a true bias tendency *it is far better to increase the number of start years, than to increase the number of ensemble members.* This is also found to be the case when the variance of the noise is doubled to represent a regional mean, rather than a global mean (not shown).

We note that the mean of the toy model realisations in the fully corrected case does not quite match the expected value (black). This is probably due to an interaction between the B_{consis} and B_{obsvar} correction terms as B_{consis} will also have a variability component, but this estimate is still the least biased.

4. Estimating the true bias in an operational decadal prediction system

S07 describe the performance of a set of hindcasts made using the HadCM3 global climate model (Gordon et al. 2000). Here we analyse a later set of ensembles, termed NoAssimPPE, which utilises the same HadCM3 GCM, but with 9 different 'perturbed physics' versions (Smith et al. 2010). These different PPE versions were chosen to sample a wide range of climate sensitivities and ENSO amplitudes (e.g. Murphy et al. 2004; Smith et al. 2010; Collins et al. 2011).

The hindcasts were initialised from model states consistent with the applied radiative forcings using start dates once per year from 1961-2001, with one 10 year prediction per model version. As in the original S07 hindcasts, the NoAssimPPE hindcasts used the 'Projection' approach to specifying external forcings (Section 2).

a. Start-time dependent forcing bias

First, we demonstrate the presence of a start-time dependent forcing bias in the NoAssimPPE hindcasts (41 start dates, 9 ensemble members, 1961-2001). Because the hindcasts use only information available at the start of the forecast, 'future' volcanic eruptions were not considered. This produces hindcasts that are biased warm when compared to observations. Also, the previous solar cycle is repeated, which is another potential source of bias.

Fig. 6 shows estimates of the natural forcings (volcanic and solar) used in the transient 20th century integrations (left) and in the prediction system (middle). The estimates for the prediction systems assume an exponential decay rate of the volcanic aerosol present at the forecast start time of 1 year and an 11 year solar cycle length. The resulting forcing bias is shown in the right column.

When integrated over all start dates an estimate of the start-time dependent forcing bias is produced (bottom right). The magnitude of the bias is dominated by the volcanic component and peaks at around $0.45\mathrm{Wm}^{-2}$ at a lead time of 3 years, subsequently dropping to around $0.30\mathrm{Wm}^{-2}$ at a lead time of 10 years.

b. Bias tendency estimates in NoAssimPPE

We now explore the expected error in the bias estimates using the results from analysis of the toy model. Fig. 7 shows the expected growth with lead time of the error in the estimated bias for NoAssimPPE (grey) where the solid (dashed) grey line indicates the expected error using 1 (9) ensemble members. The black line shows the corresponding error for the original NoAssim (S07) hindcasts (effectively 20 start dates and 16 ensemble members). The greater number of ensemble members in the original NoAssim results in a smaller expected error at short lead times (1-3 years), compared with the single member PPE system. However, the larger number of start dates in NoAssimPPE suggests a far smaller error at long lead times (5-10 years), even using a single ensemble member. The uncertainty estimates for 5-year means (horizontal grey bars) are used below in Section 5.

We next apply the bias estimate methodology developed using the toy model to annual means of global mean surface air temperature from the NoAssimPPE hindcasts (Fig. 8). We compare the hindcasts to four observational datasets (HadCRUT4 - Morice et al. 2012, GISTEMP - Hansen et al. 2010, NCEP - Kalnay et al. 1996, ERA-40 - Uppala et al. 2005), but all give consistent results. Note that the observations used are for 1961-2010, except ERA-40 which uses 1961-2001. Unless otherwise stated we use HadCRUT4 in all that follows. For the NoAssimPPE system, the raw bias tendency estimate (Fig. 8a) suggests that HadCM3 has a warm bias, which is apparently primarily due to a start-time dependent forcing bias (Fig. 8b) rather than an insufficient sampling of the observational variability (Fig. 8c). The best estimate for the true bias tendency (Fig. 8d) shows a very slight warm bias of around 0.02K/decade, which is marginally statistically significant. The interpretation of this true bias tendency is discussed in Section 5.

In addition, we note that the bias is positive over both land and sea (Fig. 8e,f). Both the spatial pattern and physical processes responsible for the bias growth will be explored in future work.

The global mean SAT bias tendency associated with the time dependent forcing error makes the largest contribution to the SAT total bias tendency (Fig. 8). Smith et al. (2007) also recognised the importance of accounting for the bias caused by volcanic eruptions. They estimated that the raw bias for NoAssim was around 0.14K/decade (consistent with Fig. 8), but they removed the forcing bias by excluding some years following volcanic eruptions. We believe that our result is more robust as we are accounting for the forcing bias more explicitly and objectively.

The lead time evolution of the ensemble mean global averaged shortwave radiation (SW) bias tendency over the ocean at the top of atmosphere (TOA) (i.e. the forcing error) using the consistent verification times is illustrated in Fig. 9a, and shows a rapid increase in downward solar radiation in the first 3-4 years to about $0.30 - 0.35\mathrm{Wm}^{-2}$ and it maintains this magnitude afterwards. This estimated forcing error and its lead time evolution are consistent with the implied surface heat flux bias tendency from vertically integrated ocean heat content (OHC) bias tendency (the implied flux bias tendency is not sensitive to the depth chosen for the integration since OHC bias tendency is mostly confined in the top 500 metres) as shown in Fig. 9b and it is also consistent with the directly estimated forcing error associated with volcanic eruptions (Fig. 9c, smoothed from Fig. 6). A caveat with using the 1961-2001 start dates for validation is that the Agung volcano in 1963 is before the consistent verification times. We have performed a sensitivity test by excluding the hindcasts from 1961, 1962, 1963, and 1964, but this does not significantly affect the results.

The relative importance of each component of the bias is illustrated in Fig. 10, which confirms that the lead time dependent forcing bias dominates. For NoAssimPPE the sampling correction (orange) is very small for global mean temperature because the number of hindcast starts dates is large. Note, however, that this contribution is expected to be larger for other variables and smaller regions. These results illustrate clearly the importance of decomposing the bias into its different components before interpreting its meaning. Furthermore, if a bias correction were to be applied to a *forecast* (rather than a hindcast), we suggest it is the underlying true bias tendency that should be used, rather than the raw bias tendency derived from the hindcasts, in contrast to some current practices (e.g. Smith et al. 2013). We plan to explore the issues surrounding the application of bias corrections to forecasts in future work.

5. Interpretation of the true bias tendency

a. Role of ocean heat uptake in bias tendency

The true bias tendency could arise either from start-time independent errors in the forcings applied to the model (e.g. errors in the specification of anthropogenic aerosols) or from errors in the transient sensitivity of the model to such forcings (or both). Errors in the transient sensitivity could themselves arise from errors in the representation of atmospheric or surface feedbacks and/or from errors in the representation of ocean heat uptake (e.g. Raper et al. 2002, Gregory and Forster 2008, Boé et al. 2009). This last factor can be examined by considering the bias tendency for global mean ocean heat content (OHC; Figure 11). As for surface air temperature the total bias is dominated by the forcing bias. The true bias tendency for the surface or top 100m is again positive, and is near zero below a few hundred metres. If insufficient ocean heat uptake were the cause of the warming bias at the surface we would expect to see a cooling bias subsurface. The fact that we don't see such a feature suggests that ocean heat uptake is not the reason for the warming bias in surface air temperature.

Further insights into the true bias tendency may be obtained by considering the biases associated with individual model versions (as distinct from the ensemble mean considered previously). Figure 12 shows that, within the PPE ensemble, there is a high positive correlation between the true bias tendency for OHC and that for SAT. This correlation again implies that variations in ocean heat uptake are not the primary cause of variations in SAT bias in NoAssim PPE.

b. Relating climate sensitivity, forcing trends and bias tendency

Next we consider the possible causes of the different true bias tendencies in the various PPE versions.

The first possible explanation is that the true bias tendency is directly related to the climate sensitivity of the model version (Figure 13a). Values for the Transient Climate Response (TCR) were obtained for each model version through separate specific experiments carried out at the UK Met Office. The HadCM3 NoAssim PPE model versions have a TCR range of 1.6-2.7K with a mean of 2.1K, which may be compared with the likely range of 1.0-2.5K from IPCC AR5 (Stocker et al. 2013). Figure 13a shows a linear relationship between the true bias tendency for global mean SAT and TCR, in which the most sensitive models give the largest warming bias tendency, with a correlation coefficient of 0.89. This high correlation suggests that the true bias tendency may be providing very useful information about the sensitivity of the underlying model. The correlation between TCR and the uncorrected bias tendency is 0.75, so the corrections have also improved this relationship. In addition, since a perfect model should yield a true bias tendency of zero, we can use this relationship to estimate a likely range for TCR.

A Monte-Carlo approach is used to fit regression lines to the data by perturbing the true bias tendency of each model version, taking into account the bias tendency uncertainty (0.016K, calculated from the toy model). The distribution of the intercepts of these lines with the $y = 0$ line (corresponding to zero true bias tendency) then provides an observationally constrained range for TCR. We find that the $5-95\%$ range for TCR constrained in this way is 1.4-1.8K with a median of 1.6K using HadCRUT4 (Figure 13c). This range is considerably narrower than the corresponding likely range from IPCC AR5 of 1.0-2.5K, and observation-based ranges of 1.3-2.3K (Gregory and Forster 2008) and 0.9-2.0K (Otto et al. 2013). With doubled estimates for the uncertainty in the true bias tendency the range from this study becomes 0.9-1.9K. The standard version of HadCM3 has a TCR of 2.0K (Randall et al. 2007).

The constrained ranges of TCR for different observational data sets, are summarized in Table 1. Results indicate that the median and the ranges of the constrained TCR are only slightly sensitive to the data that is used to validate the hindcasts, with the other datasets producing values of TCR about 0.15K higher. The reduced spread of TCR is a robust feature and so the underlying SAT true bias tendency from the decadal climate hindcasts could be used to constrain the model TCR, complementing other approaches proposed in the literature (e.g. Allen et al. 2000, Stott and Forest 2007, Gregory and Forster 2008, Knutti and Tomassini 2008, Murphy 2010, Tett et al. 2013). It is also interesting to note that having a range of models with widely different TCR has proved very useful in this analysis, especially to constrain the upper end of our TCR ranges.

However, there is another possible explanation for the true bias tendency differences. When considering the role of TCR we have assumed that the forcing trends in each PPE version are the same. However, Harris et al. (2013) recently demonstrated that the different PPE versions of HadCM3 have different non-greenhouse gas (GHG) forcing, likely due to the different interaction of aerosols with low clouds. The relationship is such that versions of HadCM3 with a low TCR, and negative bias tendency, also have a cooling trend from non-GHG forcing from 1961-2010, and this could potentially contribute to the relationship between TCR and true bias tendency.

Figure 13b relates the true bias tendency to the non-GHG forcing trends for the different PPE model versions. The forcing data are taken from Harris et al. (2013), and linear trends have been fitted from 1961-2010, excluding years with, and shortly after, volcanic eruptions. This provides an estimate of the non-GHG forcing trends and the observed relationship can be used to produce an improved constraint on the non-GHG forcing trend, which is found to be negative, unlike in the majority of the model versions.

There are therefore two possible causes for the relationship between perturbed parameter versions of HadCM3 and the true bias tendency, i.e. it is clear that the parameter perturbations affect both the TCR and the non-GHG forcing trends and that both factors influence the true bias tendency. Trying to separate the two effects is beyond the scope of this paper, but further work will use the spatial patterns, and other climate variables, to further understand the causes of the bias tendencies. However, we note that if both factors are playing a role then the constrained ranges for TCR and non-GHG forcing would broaden.

An additional related caveat is that if there is a *systematic* error (i.e. common to all model versions) in the trends in the radiative forcing applied to the model then this would also affect the true bias tendency. For example, if the forcing trends were systematically too large then the true bias tendency would also be too large, and vice versa. The result of any such bias would be to displace all the data in Figures 13a,b vertically along the true bias tendency axis. Such a displacement would shift the constrained ranges but would not broaden the distributions. This caveat should be kept in mind when interpreting our results.

One possible approach to addressing these various caveats would be a multi-model study where the forcings are likely to be different for each model, and this is planned further work.

6. Conclusions and discussion

We have explored the estimation of bias in a toy model of a decadal prediction system, and applied the techniques developed to analyse the bias of operational predictions of global mean temperature. We have focused on hindcasts initialised from model states, rather than from observations, and examined the bias tendency in particular. The main findings can be summarised as follows:

- The total bias tendency can be separated into several components, namely: a contribution from sampling uncertainty due to internal variability, a start-time dependent forcing bias tendency, and the true bias tendency.

- We have shown how the contributions from sampling uncertainty and start-time dependent forcing bias can be estimated, and removed, to give a better (lower variance and less biased) estimate of the true bias tendency. We argue that it is the true bias tendency, not the total bias tendency, that should be used to adjust decadal forecasts.

- The true bias tendency is attributable to: 1) errors in the sensitivity of the underlying model to forcing, and/or 2) start-time independent errors in the specification of forcing (e.g. errors in the specification of anthropogenic aerosols).

- To improve estimates of bias tendencies, more hindcast start dates are more beneficial than more ensemble members.

- The UK Met Office NoAssim PPE prediction system exhibits, in the ensemble mean, a small positive true bias tendency in hindcasts of global mean surface air temperature, and this is marginally statistically significant. We have demonstrated that this bias is not attributable to insufficient ocean heat uptake.

- The different true bias tendencies in global mean surface air temperature in the various PPE versions can be used to constrain relevant physical properties of the models, such as the TCR and non-GHG forcing trends.

There are a number of caveats to the findings above. In the toy model, we have assumed linear trends. However, we do not believe that this compromises the decomposition of the bias tendency into its different terms. Secondly, we assumed that the toy model has the same variability properties as the toy observations. This is unlikely to hold perfectly in an operational setting as there is a broad spread in simulated variability amongst different models (Hawkins and Sutton 2012) and even amongst the different PPE versions of HadCM3 (Ho et al. 2013), but this would only change the number of start dates and ensemble members required to reliably estimate the bias. Most importantly, we have assumed the radiative forcings imposed in the decadal hindcasts are correct, as discussed in Section 5.

In the decadal hindcast experiments for CMIP5, the standard start dates are every 5 years (Meehl et al. 2009; Taylor et al. 2012). In this situation there is no way of estimating the consistent bias on annual timescales. Therefore, any lead time dependent errors in the forcing cannot be removed. However, in the 'Tier 1' CMIP5 predictions, the complete volcanic and solar forcings are assumed known, so there should be little start-time dependent forcing bias. In other suggested experiments this is not the case. We suggest that the design of future decadal prediction experiments should consider start dates every year to allow for any start-time dependent forcing bias to be removed.

We believe that the analysis of bias tendencies has considerable potential to provide further insights into climate models and the real climate system. We note that Masson and Knutti (2013) suggest that perturbed-physics ensembles and multi-model ensembles can behave differently and show opposite emergent constraints so it would be valuable to repeat this analysis using a wider range of operational prediction systems.

Beyond the global means considered in this paper there is a great deal of information in the spatial patterns of bias growth for a range of variables, and we have begun work to analyse these patterns. Lastly, there is an obvious need to examine how the growth of biases in a system initialised from model states is related to the growth of biases in a system initialised from observational states. This work involves many challenges but is essential for the development of decadal predictions.

Acknowledgments.

We thank Glen Harris for providing the forcing data from his important study, and for valuable discussions. We also thank two anonymous reviewers for their helpful comments which improved the manuscript. The research leading to this paper has received support from NCAS-Climate (EH, BD and RS), from the European Community's 7th framework programme (FP7) under grant agreement No. GA212643 (THOR) (EH, DS) and from the UK NERC funded EQUIP (EH) and VALOR (JR) projects. DS was also supported by the joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101), and the EU FP7 COMBINE project.

REFERENCES

Allen, M. R., P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, **407**, 617–620, doi: 10.1038/35036559.

Boé, J., A. Hall, and X. Qu, 2009: Deep ocean heat uptake as a major source of spread in transient climate change simulations. *Geophysical Research Letters*, **36**, L22 701, doi: 10.1029/2009GL040845.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J. Geophys. Res.*, **111**, D12106, doi:10.1029/2005JD006548.

Collins, M., B. B. B. Booth, B. Bhaskaran, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb, 2011: Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Clim. Dyn.*, **36**, 1737–1766, doi:10.1007/s00382-010-0808-0.

Ferranti, L. and P. Viterbo, 2006: The european summer of 2003: Sensitivity to soil water initial conditions. *J. Climate*, **19**, 3659–3680, doi:10.1175/JCLI3810.1.

Goddard, L., et al., 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, **40**, 245–272, doi:10.1007/s00382-012-1481-2.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.*, **16**, 147–168.

Gregory, J. M. and P. M. Forster, 2008: Transient climate response estimated from radiative forcing and observed temperature change. *Journal of Geophysical Research: Atmospheres*, **113**, D23 105, doi:10.1029/2008JD010405.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, doi:10.1029/2010RG000345.

Harris, G. R., D. M. Sexton, B. B. Booth, M. Collins, and J. M. Murphy, 2013: Probabilistic projections of transient climate change. *Climate Dynamics*, **40**, 2937–2972, doi:10.1007/s00382-012-1647-y.

Hawkins, E. and R. Sutton, 2012: Time of emergence of climate signals. *Geophysical Research Letters*, **39**, L01 702, doi:10.1029/2011GL050087.

Ho, C. K., E. Hawkins, L. Shaffrey, J. Bröcker, L. Hermanson, J. M. Murphy, D. M. Smith, and R. Eade, 2013: xamining reliability of seasonal to decadal sea surface temperature forecasts: the role of ensemble dispersion. *Geophys. Res. Lett.*, **in press**, doi:10.1002/2013GL057630.

Kalnay, E., et al., 1996: The NCEP/NCAR 40-year reanalysis project. *BAMS*, **77**, 437–471.

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453**, 84–88, doi:10.1038/nature06921.

Knutti, R. and L. Tomassini, 2008: Constraints on the transient climate response from observed global temperature and ocean heat uptake. *Geophysical Research Letters*, **35**, L09 701, doi:10.1029/2007GL032904.

Masson, D. and R. Knutti, 2013: Predictor screening, calibration, and observational constraints in climate model ensembles: an illustration using climate sensitivity. *J. Climate*, **26**, 887–898, doi:10.1175/JCLI-D-11-00540.1.

Meehl, G. A., L. Goddard, G. Boer, R. Burgman, and et al., 2013: Decadal climate prediction: An update from the trenches. *BAMS*, in press, doi:10.1175/BAMS-D-12-00241.1.

Meehl, G. A., et al., 2009: Decadal prediction: can it be skillful? *Bull. Amer. Met. Soc.*, **90**, 1467–1485, doi:10.1175/2009BAMS2607.1.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophy. Res. Atmos.*, **117**, D08 101, doi:10.1029/2011JD017187.

Murphy, D. M., 2010: Constraining climate sensitivity with linear fits to outgoing radiation. *Geophysical Research Letters*, **37**, L09 704, doi:10.1029/2010GL042911.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:10.1038/nature02771.

Oreskes, N., D. A. Stainforth, and L. A. Smith, 2010: Adaptation to global warming: Do climate models tell us what we need to know? *Philosophy of Science*, **77 (5)**, 1012–1028, doi:10.1086/657428.

Otto, A., F. Otto, O. Boucher, and et al., 2013: Energy budget constraints on climate response. *Nature Geosci.*, **6**, 415–416, doi:10.1038/ngeo1836.

Pohlmann, H., J. Jungclaus, A. Kohl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate*, **22**, 3926–3938, doi:10.1175/2009JCLI2535.1.

Randall, D. A., et al., 2007: *Climate Models and Their Evaluation. In: Climate Change 2007: The Physical Science Basis.* Cambridge University Press, Cambridge, UK.

Raper, S. C. B., J. M. Gregory, and R. J. Stouffer, 2002: The role of climate sensitivity and ocean heat uptake on aogcm transient temperature response. *J. Climate*, **15**, 124–130, doi:10.1175/1520-0442(2002)015⟨0124:TROCSA⟩2.0.CO;2.

Robson, J. I., 2010: Understanding the performance of a decadal prediction system. Ph.D. thesis, University of Reading, UK.

Robson, J. I., R. T. Sutton, and D. M. Smith, 2012: Initialized decadal predictions of the rapid warming of the north atlantic ocean in the mid 1990s. *Geophysical Research Letters*, **39**, L19 713, doi:10.1029/2012GL053370.

Smith, D. M., S. Cusack, A. W. Colman, C. K. Folland, G. R. Harris, and J. M. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796–799, doi:10.1126/science.1139540.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nature Geosci.*, **3**, 846–849, doi:10.1038/ngeo1004.

Smith, D. M., et al., 2013: Real-time multi-model decadal climate predictions. *Climate Dynamics*, in press, doi:10.1007/s00382-012-1600-0.

Stocker, T. F., et al., (Eds.) , 2013: *IPCC, 2013: Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK.

Stott, P. A. and C. E. Forest, 2007: Review. Ensemble climate predictions using climate models and observational constraints. *Phil. Trans. R. Soc. A*, **365**, 2029–2052, doi:10.1098/rsta.2007.2075.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tett, S., D. Rowlands, M. Mineter, and C. Cartis, 2013: Can top of atmosphere radiation measurements constrain climate predictions? part 2: Climate sensitivity. *J. Climate*, in press.

Toniazzo, T. and S. Woolnough, 2013: Development of warm SST errors in the southern tropical Atlantic in CMIP5 decadal hindcasts. *Clim. Dyn.*, in press, doi:10.1007/s00382-013-1691-2.

Uppala, S. M., et al., 2005: The ERA-40 re-analysis. *QJRMS*, **131**, 2961–3012, doi:10.1256/qj.04.176.

van Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger, 2012: Decadal prediction skill in a multi-model ensemble. *Climate Dynamics*, **38**, 1263–1280, doi:10.1007/s00382-012-1313-4.

Yeager, S., A. Karspeck, G. Danabasoglu, J. Tribbia, and H. Teng, 2012: A decadal prediction case study: Late 20th century north atlantic ocean heat content. *J. Climate*, **25**, 5173–5189.

List of Tables

1 The 5-95% ranges and medians (in brackets) of the original TCR (K) and the bias constrained values using a Monte-Carlo approach of linear fits to TCR against different observations. 22

TABLE 1. The 5-95% ranges and medians (in brackets) of the original TCR (K) and the bias constrained values using a Monte-Carlo approach of linear fits to TCR against different observations.

	TCR
Original	1.61-2.64 (2.17)
Constrained ranges	
ERA40	1.65-1.99 (1.82)
NCEP	1.59-1.91 (1.75)
GISS	1.61-1.93 (1.77)
HadCRUT4	1.45-1.83 (1.64)

List of Figures

1 A schematic illustrating the definition of B_{obsvar} (Eqn. 7) and consistent verification times (Section 2e). (a-c) Black lines show pseudo-observations, the red lines show pseudo-predictions (with noise) for three lead times (τ) as labelled, and the grey regions indicate the area integrated in the definition of B_{obsvar}. The blue bars indicate the range of times which are considered 'consistent', i.e. where all lead times can be simultaneously assessed. (d) B_{obsvar} for all verification times (black) and consistent verification times (blue). (e) Same as (d) for B'_{obsvar}. 27

2 (a) Example of a simple pseudo-prediction system, including observations (black), predictions (red), the true forced trend (solid blue) and estimated forced trend (dashed blue). (b) The bias tendency estimates for the predictions in (a), showing the true bias tendency (dark grey), the raw bias tendency estimate (solid blue), the bias tendency corrected using observed variability (B_{obsvar}) for the cases when the forced trend is known (black) and unknown (dashed blue). (c) 10 examples of the bias estimates in (b) with different realisations of the observations. 28

3 The spread in 1000 realisations of the bias tendency estimates, an example of which is shown in Fig. 2, for the raw bias tendency (solid black) and corrected bias tendency (dashed black) at a lead time of 10 years. The magnitude of the true bias is shown in grey, indicating that, for this choice of toy model parameters, the bias could be detected with $L \approx 16$ (20) hindcast start dates if the correction is made (or not). 29

4 (a) Example of a pseudo-prediction system with a lead time dependent bias, including observations (black), hindcasts (red), the true forced trend (solid blue) and estimated forced trend (dashed blue), including a mock volcanic eruption. (b) The bias tendency estimates for the predictions in (a), showing the true bias tendency (dark grey), true forcing bias tendency (light grey), the raw bias tendency estimates (blue), the bias tendency using consistent verification times (red) and the bias tendency estimates corrected using the consistent bias tendency (green). The dashed blue and green lines are corrected using B'_{obsvar}. 30

5 Spread in bias tendency estimates at a lead time of 10 years, as a function of the number of ensemble members considered, for (a) fully corrected bias estimate, (b) no observed variability correction, (c) no lead time dependent forcing bias correction, and (d) the raw bias. 31

6 An estimate of the start-time dependent forcing bias in the NoAssim prediction system (Smith et al. 2010). The left column shows the forcing estimates used in the transient integrations, the middle column shows the estimated forcing used in NoAssimPPE, and the right column shows the difference. The eruptions of Agung, El Chichon and Pinatubo are the main cause of the bias. 32

7 Toy model estimates for the error in true bias tendency estimates for the hindcast setup of two operational prediction systems, namely NoAssim1 (Smith et al. 2007) and NoAssim PPE (Smith et al. 2010). NoAssim1 uses 20 years of hindcasts, with an effective ensemble size of 16 members (black line). NoAssim PPE uses 40 years of hindcasts with 9 different perturbed physics versions of the model, each with a single member. These can be considered as independent single member ensembles (solid grey) or as a 9-member ensemble (dashed grey). The horizontal error bars indicate the errors for 5 year mean predictions for NoAssimPPE (single members). 33

8 Bias tendency estimates (K) for global mean surface air temperature using NoAssim PPE. Different colours represent different observational datasets. (a) Raw bias. (b) Consistent verification times bias which is an estimate of the start-time dependent forcing bias. (c) Raw bias corrected by obsvar. (d) The true bias estimate, which is (c)-(b). The error ranges in (d) are derived from the toy model (Fig. 7) and are shown relative to the ERA-40 results. 34

9 Time evolution of ensemble mean (a) true bias tendency (Wm^{-2}) in shortwave radiation at the top of atmosphere (TOA) of HadCM3 NoAssim PPE hindcasts for the period 1961-2001 against ERA-40 data set, (b) implied surface heat flux bias tendency (Wm^{-2}) from integrated ocean heat content (OHC) bias for top 1500 metres against the Met Office ocean analysis and (c) estimated global mean error (Wm^{-2}) associated with volcanic forcing in hindcasts. 35

10 The components of the total bias tendency for NoAssimPPE against HadCRUT4 data. The total bias tendency (black) is dominated by the lead-time dependent forcing bias (green). The magnitude of the forcing bias is qualitatively consistent with the magnitude of the forcing errors (Fig. 6). 36

11 Time evolutions of ensemble mean bias tendencies (K) for ocean temperature at 5m and ocean heat content (top 100m and top 500m) of HadCM3 NoAssim PPE hindcasts for the period 1961-2010 against Met Office ocean analysis data. (a) using all verification times (1961-2010), (b), using consistent verification times (1971-2001), (c) true bias tendency with linear trend removed in the analysis before calculating bias tendency associated with observed variability. (d) Time evolution of ensemble mean true bias tendency (K) as a function of depth for global ocean temperature for HadCM3 NoAssim PPE hindcasts for the period 1961-2010 against the Met Office ocean analysis. 37

12 Relationships between global mean SAT true bias tendencies (K) (against HadCRUT4 data) and global mean OHC (top 1000m) bias tendencies (against the Met Office ocean analysis) for 9 PPE model versions. (a) average for lead years 1-5, and (b) average for lead years 6-10. 38

13 Relationships between the lead years 6-10 averaged global mean SAT true bias tendencies (K) against HadCRUT4 data for each version of PPE hindcasts for (a) TCR and (b) non-GHG aerosol forcing trend, using 9 PPE model versions. The error bars for bias tendency are based on the toy model (Fig. 7). Grey lines are example linear fits to TCR and to the non-GHG aerosol forcing trend using a Monte-Carlo approach, and the red lines are the best fit. The constrained ranges of TCR and the non-GHG aerosol forcing trend are shown as black bars assuming a true bias tendency error of 0.016K (solid) and 0.032K (dashed). Other ranges for TCR (Stocker et al. 2013,Gregory and Forster 2008 - denoted GF08) ranges are also given. (c,d) estimated probability distribution functions (PDFs) of unconstrained (blue) and constrained (full black and dotted black) TCR and non-GHG aerosol forcing trends. The dashed black lines indicate the PDF for doubled uncertainties in the true bias tendency. 39

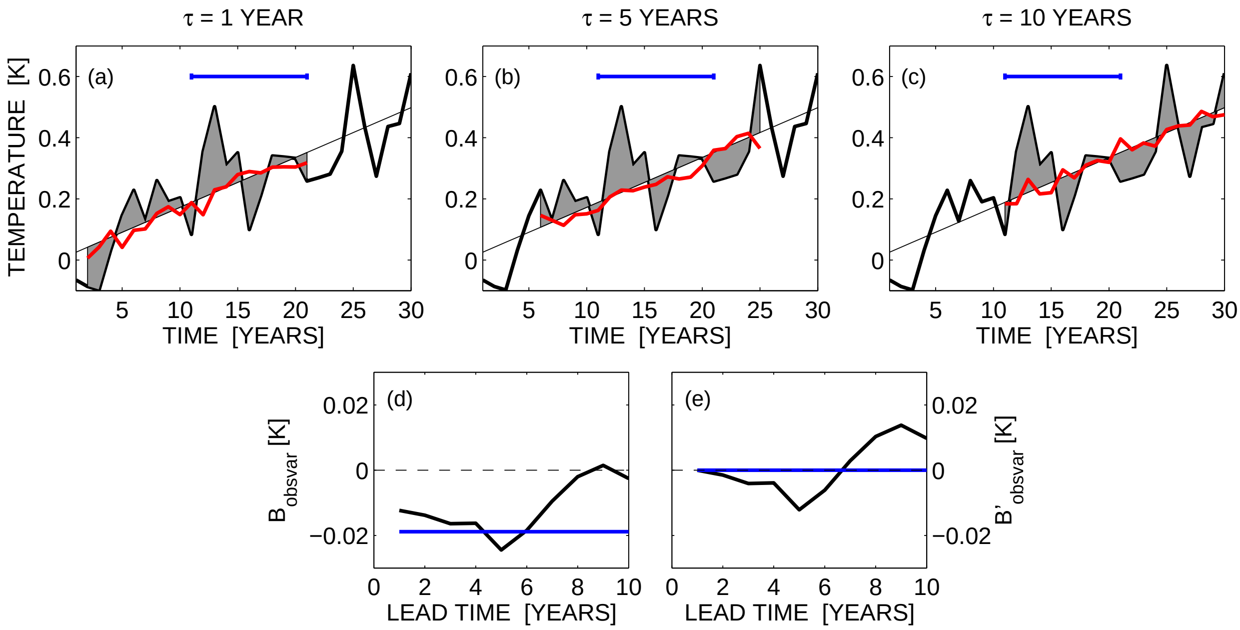

FIG. 1. A schematic illustrating the definition of B_{obsvar} (Eqn. 7) and consistent verification times (Section 2e). (a-c) Black lines show pseudo-observations, the red lines show pseudo-predictions (with noise) for three lead times (τ) as labelled, and the grey regions indicate the area integrated in the definition of B_{obsvar}. The blue bars indicate the range of times which are considered 'consistent', i.e. where all lead times can be simultaneously assessed. (d) B_{obsvar} for all verification times (black) and consistent verification times (blue). (e) Same as (d) for B'_{obsvar}.

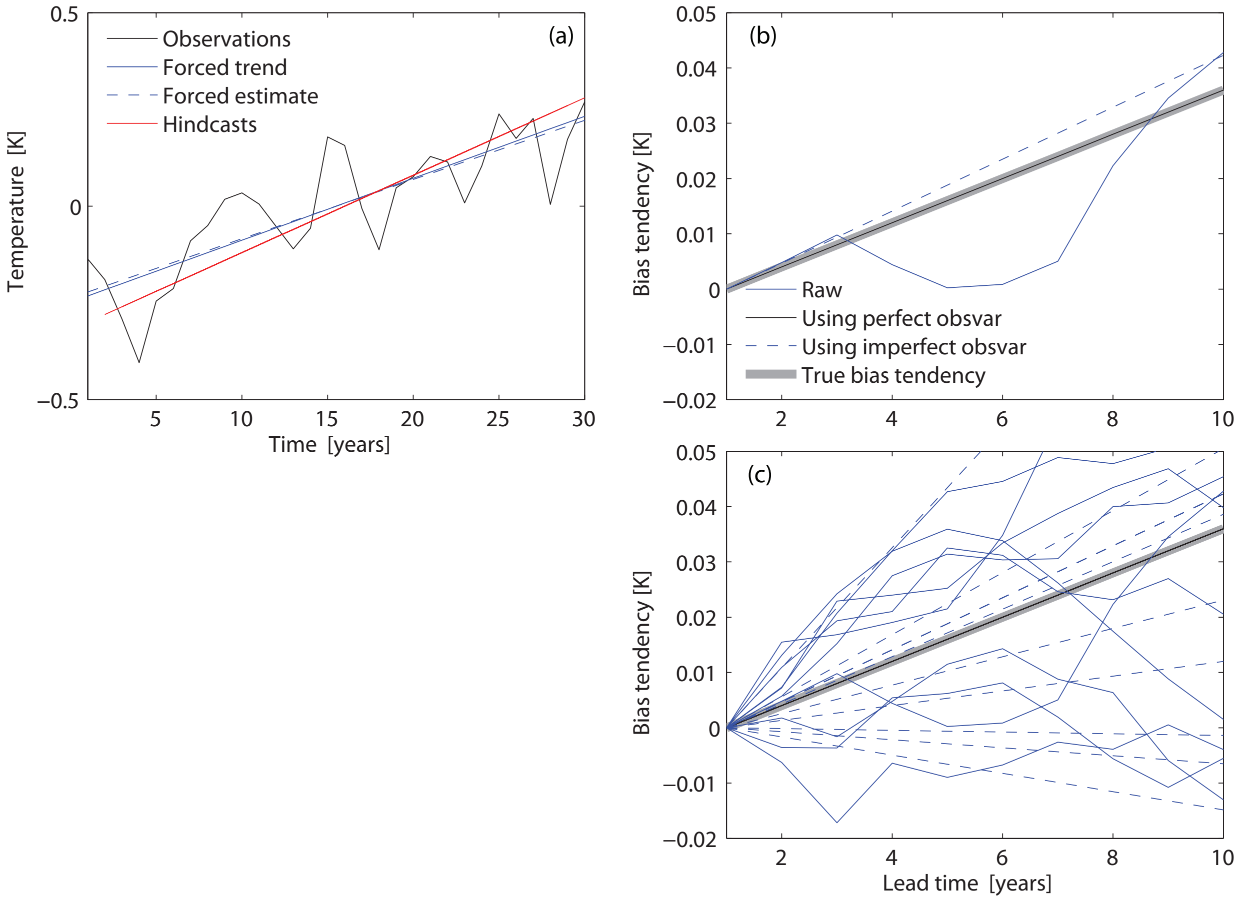

FIG. 2. (a) Example of a simple pseudo-prediction system, including observations (black), predictions (red), the true forced trend (solid blue) and estimated forced trend (dashed blue). (b) The bias tendency estimates for the predictions in (a), showing the true bias tendency (dark grey), the raw bias tendency estimate (solid blue), the bias tendency corrected using observed variability (B_{obsvar}) for the cases when the forced trend is known (black) and unknown (dashed blue). (c) 10 examples of the bias estimates in (b) with different realisations of the observations.

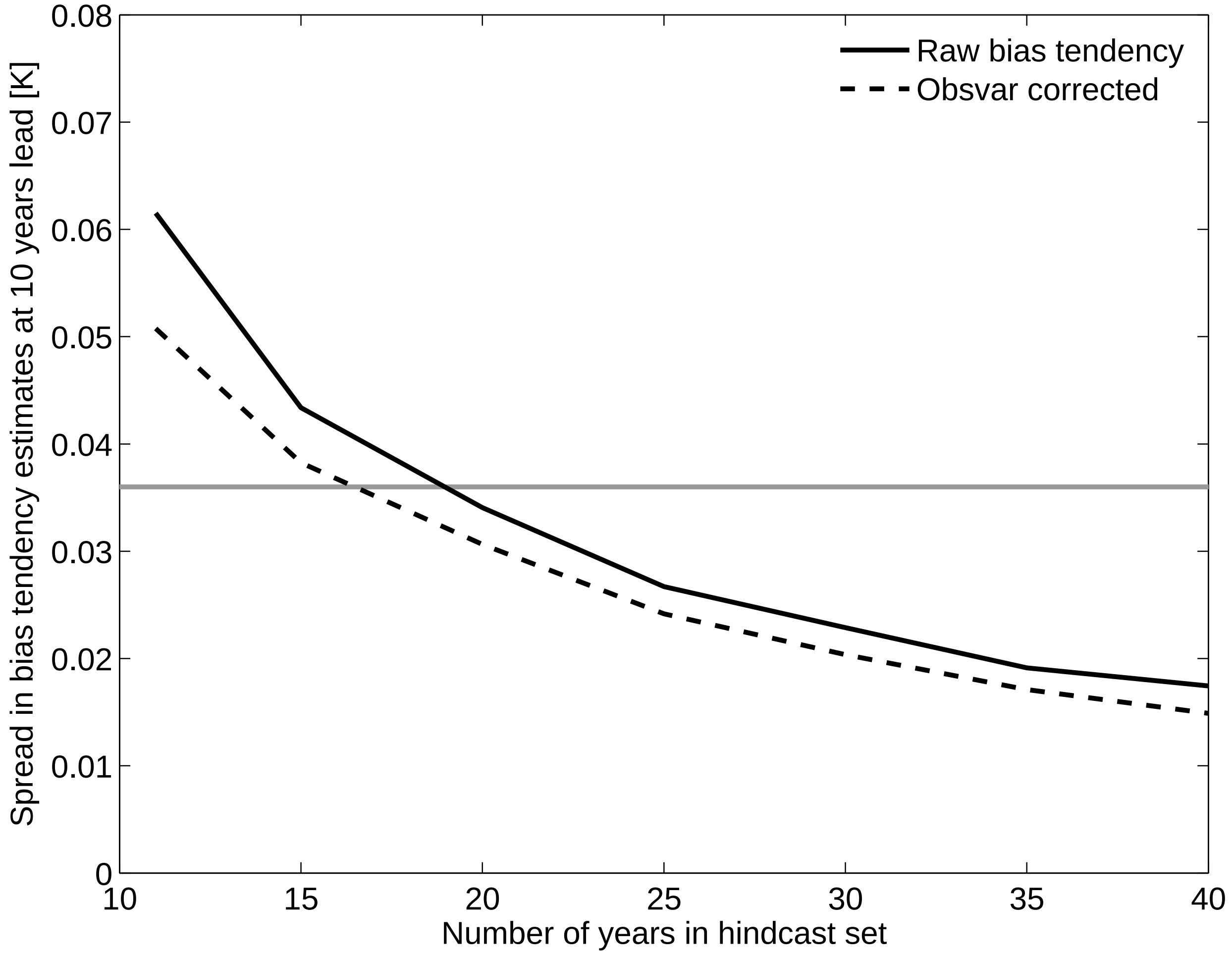

FIG. 3. The spread in 1000 realisations of the bias tendency estimates, an example of which is shown in Fig. 2, for the raw bias tendency (solid black) and corrected bias tendency (dashed black) at a lead time of 10 years. The magnitude of the true bias is shown in grey, indicating that, for this choice of toy model parameters, the bias could be detected with $L \approx 16$ (20) hindcast start dates if the correction is made (or not).

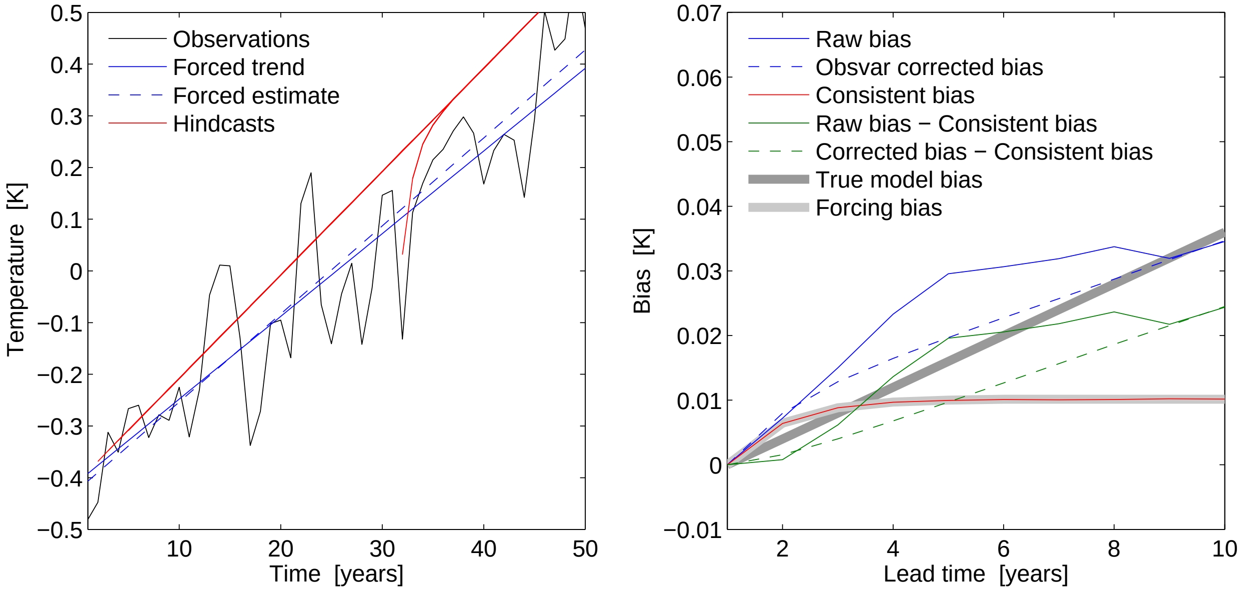

FIG. 4. (a) Example of a pseudo-prediction system with a lead time dependent bias, including observations (black), hindcasts (red), the true forced trend (solid blue) and estimated forced trend (dashed blue), including a mock volcanic eruption. (b) The bias tendency estimates for the predictions in (a), showing the true bias tendency (dark grey), true forcing bias tendency (light grey), the raw bias tendency estimates (blue), the bias tendency using consistent verification times (red) and the bias tendency estimates corrected using the consistent bias tendency (green). The dashed blue and green lines are corrected using B'_{obsvar}.

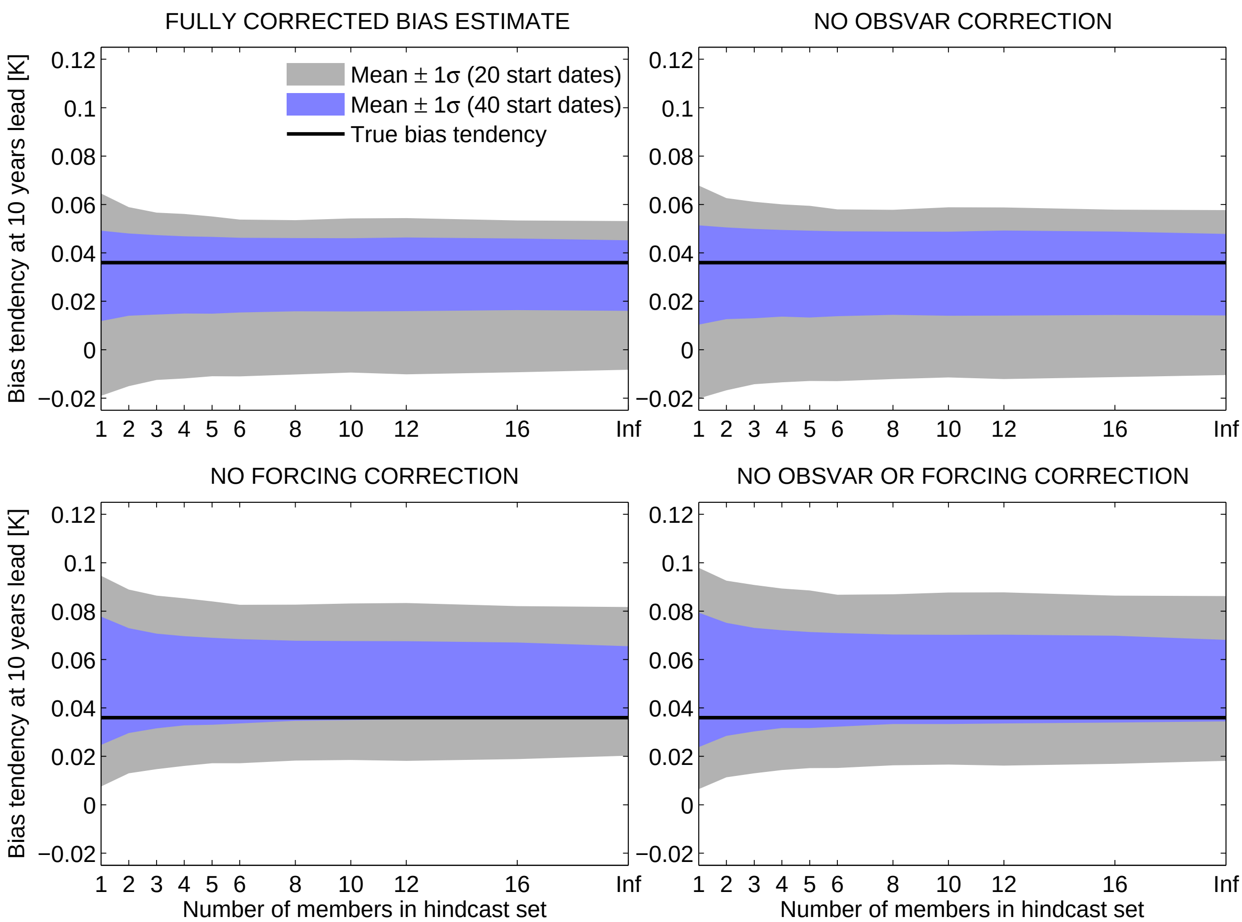

FIG. 5. Spread in bias tendency estimates at a lead time of 10 years, as a function of the number of ensemble members considered, for (a) fully corrected bias estimate, (b) no observed variability correction, (c) no lead time dependent forcing bias correction, and (d) the raw bias.

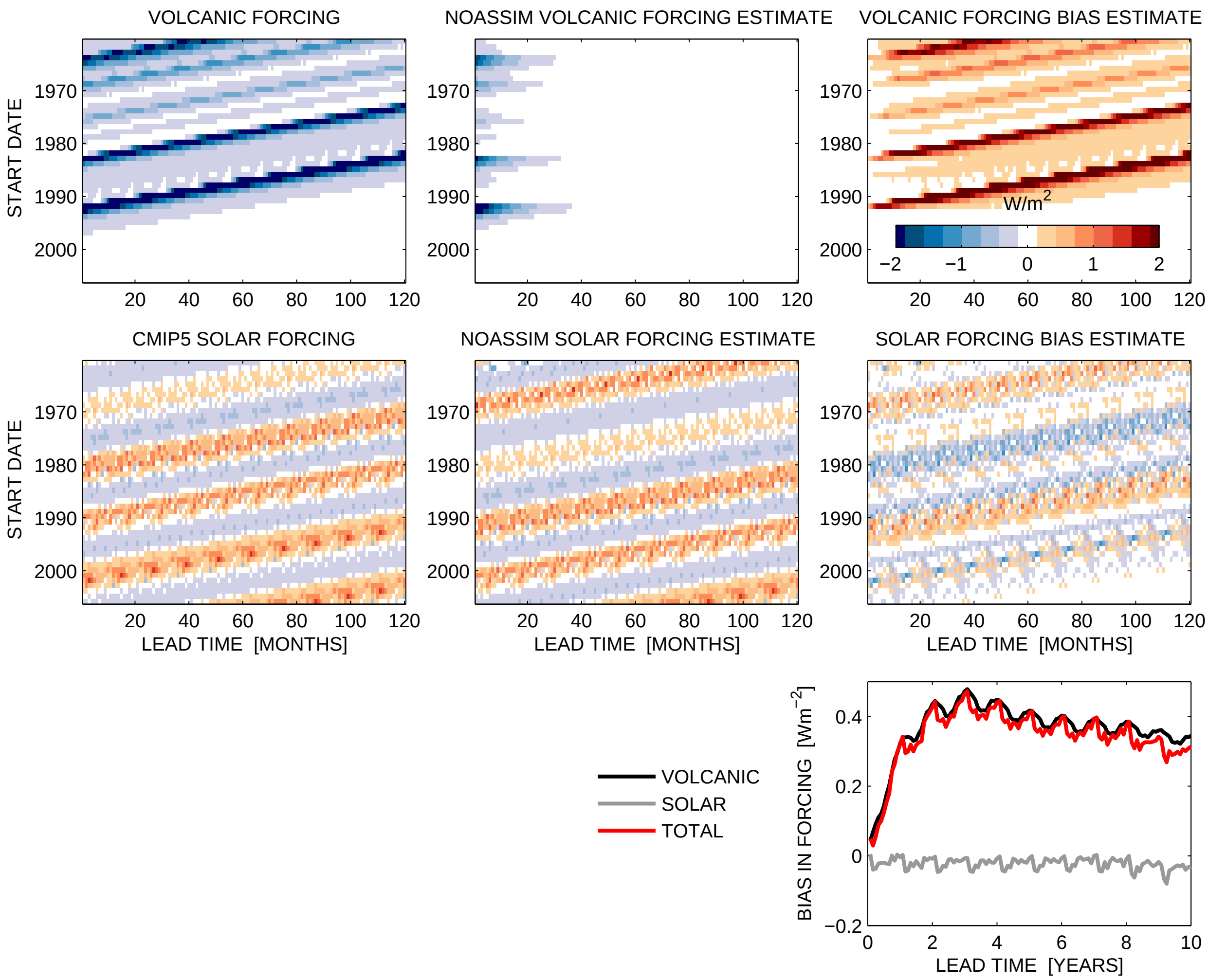

FIG. 6. An estimate of the start-time dependent forcing bias in the NoAssim prediction system (Smith et al. 2010). The left column shows the forcing estimates used in the transient integrations, the middle column shows the estimated forcing used in NoAssimPPE, and the right column shows the difference. The eruptions of Agung, El Chichon and Pinatubo are the main cause of the bias.

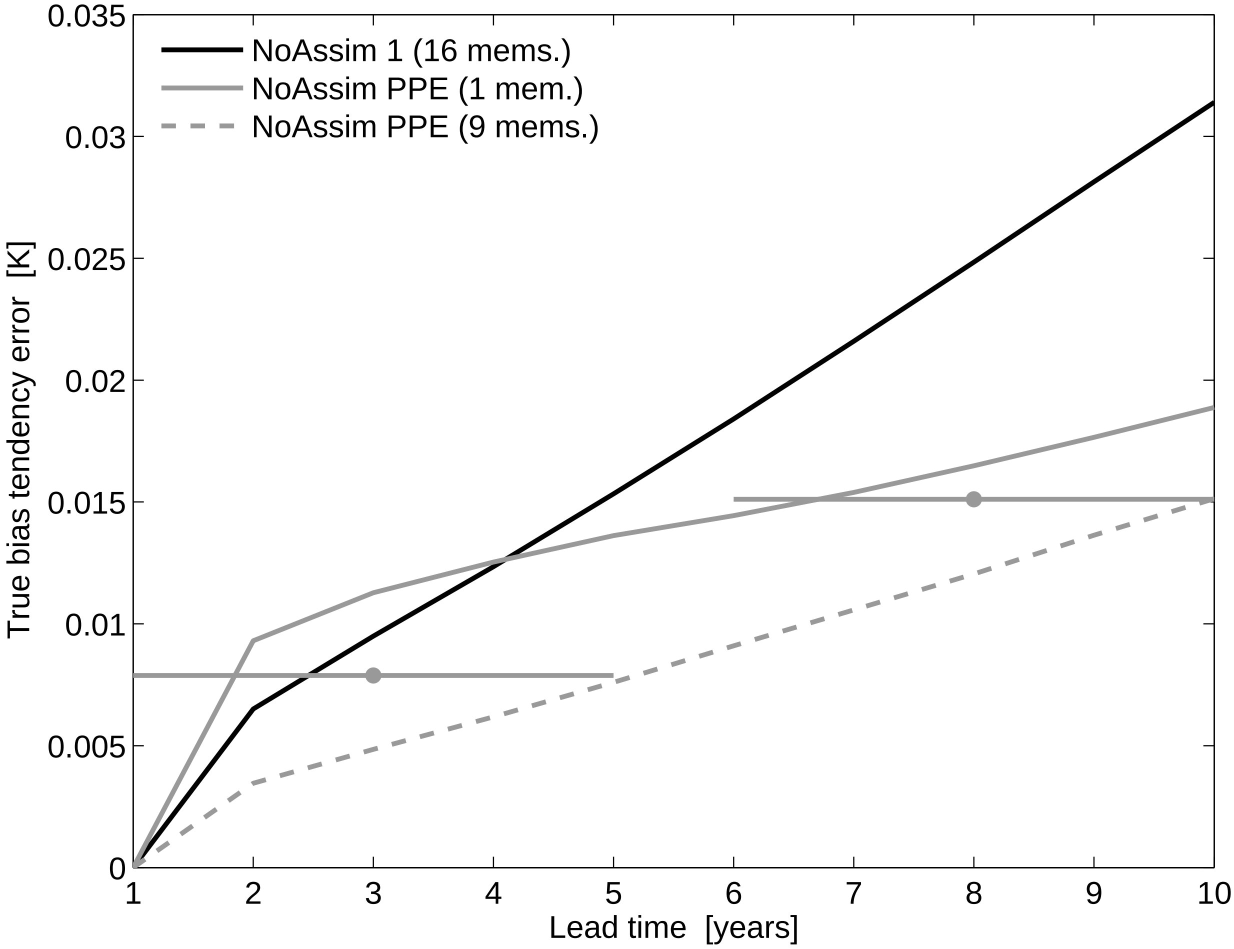

FIG. 7. Toy model estimates for the error in true bias tendency estimates for the hindcast setup of two operational prediction systems, namely NoAssim1 (Smith et al. 2007) and NoAssim PPE (Smith et al. 2010). NoAssim1 uses 20 years of hindcasts, with an effective ensemble size of 16 members (black line). NoAssim PPE uses 40 years of hindcasts with 9 different perturbed physics versions of the model, each with a single member. These can be considered as independent single member ensembles (solid grey) or as a 9-member ensemble (dashed grey). The horizontal error bars indicate the errors for 5 year mean predictions for NoAssimPPE (single members).

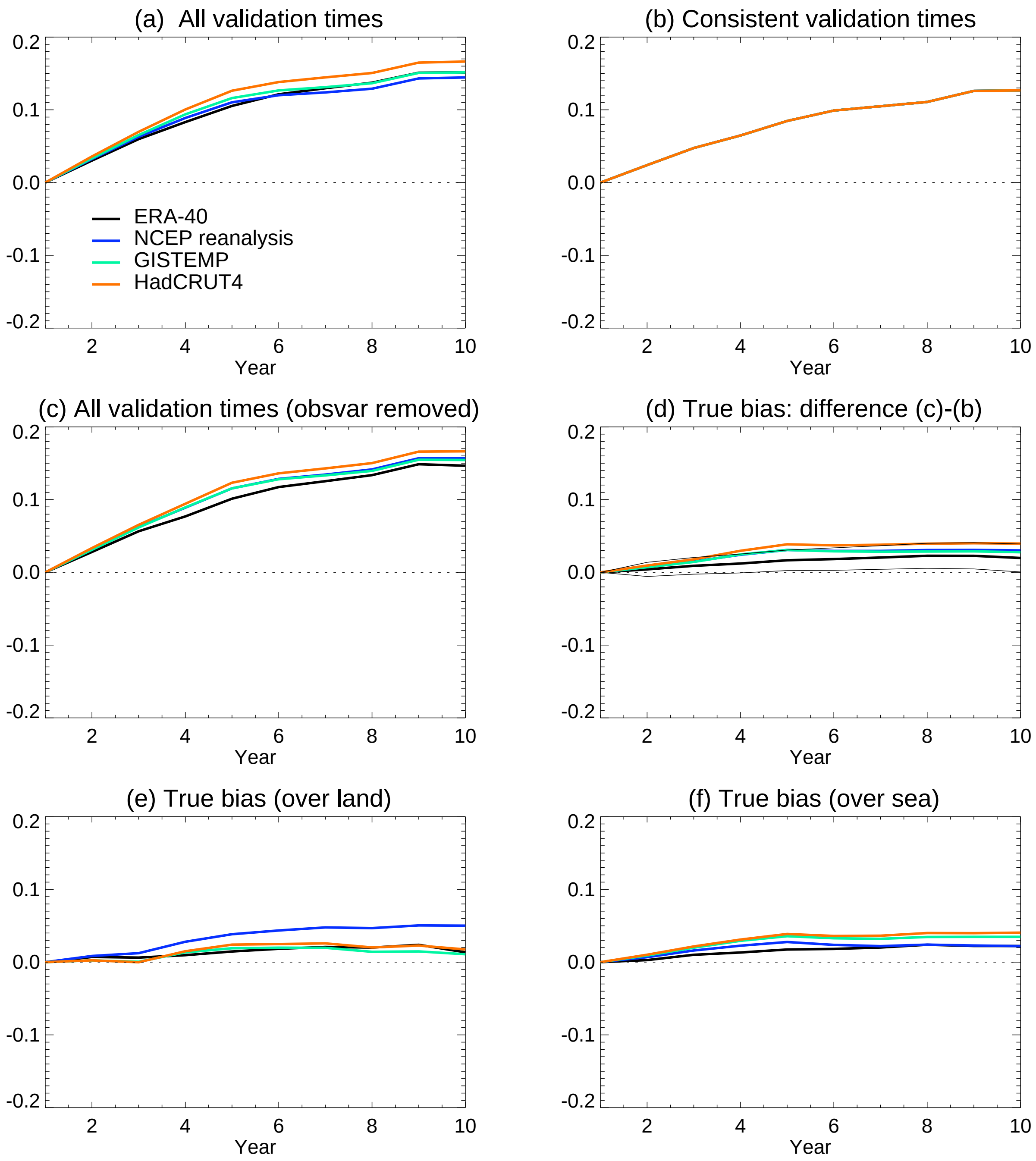

FIG. 8. Bias tendency estimates (K) for global mean surface air temperature using NoAssim PPE. Different colours represent different observational datasets. (a) Raw bias. (b) Consistent verification times bias which is an estimate of the start-time dependent forcing bias. (c) Raw bias corrected by obsvar. (d) The true bias estimate, which is (c)-(b). The error ranges in (d) are derived from the toy model (Fig. 7) and are shown relative to the ERA-40 results.

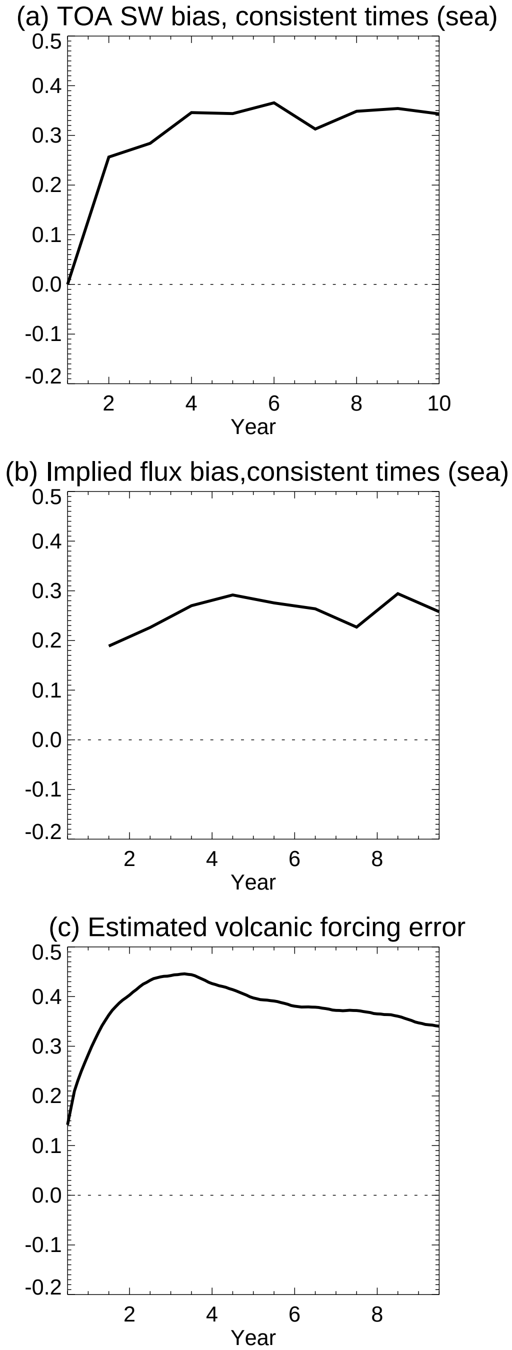

FIG. 9. Time evolution of ensemble mean (a) true bias tendency (Wm^{-2}) in shortwave radiation at the top of atmosphere (TOA) of HadCM3 NoAssim PPE hindcasts for the period 1961-2001 against ERA-40 data set, (b) implied surface heat flux bias tendency (Wm^{-2}) from integrated ocean heat content (OHC) bias for top 1500 metres against the Met Office ocean analysis and (c) estimated global mean error (Wm^{-2}) associated with volcanic forcing in hindcasts.

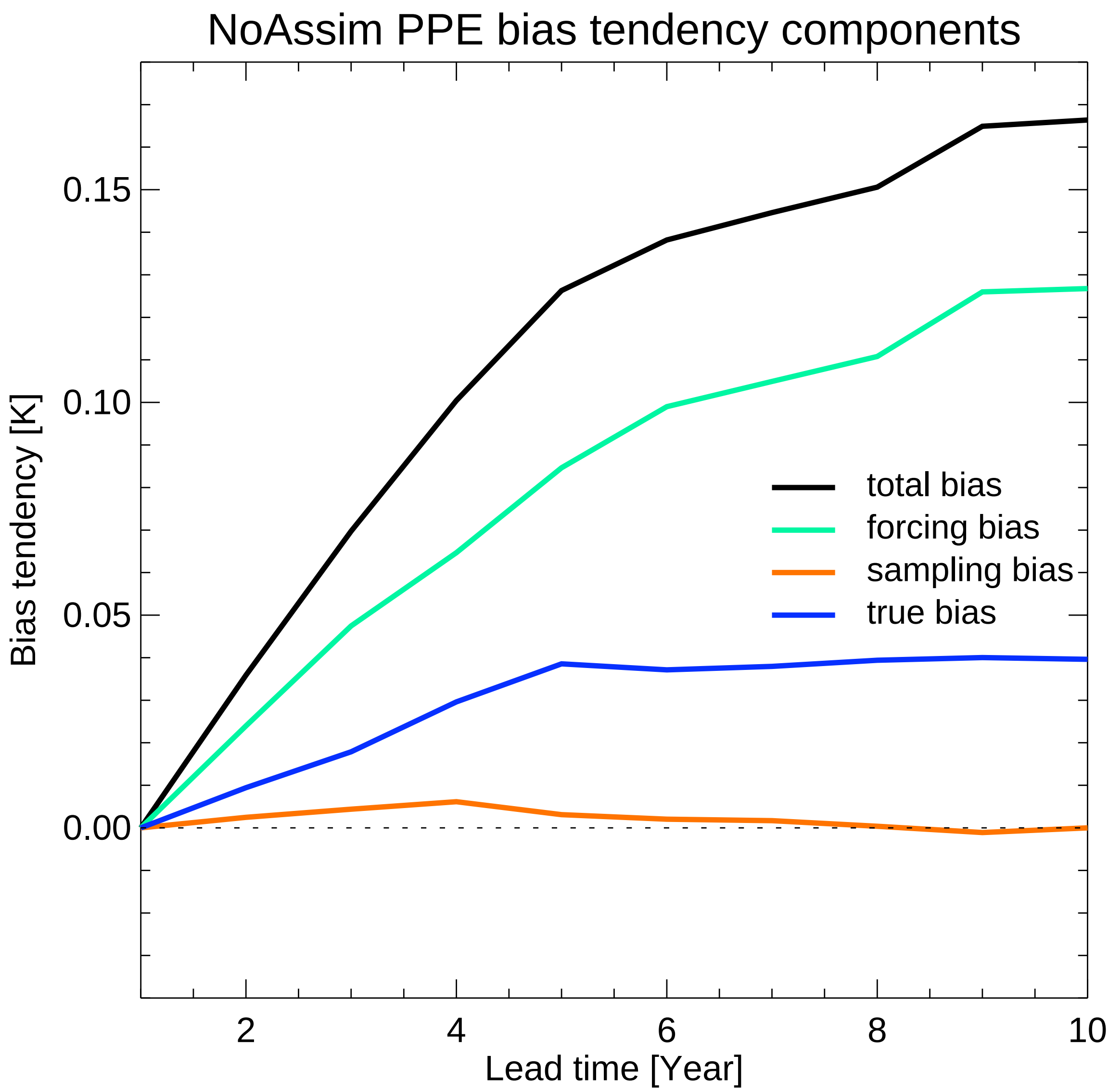

FIG. 10. The components of the total bias tendency for NoAssimPPE against HadCRUT4 data. The total bias tendency (black) is dominated by the lead-time dependent forcing bias (green). The magnitude of the forcing bias is qualitatively consistent with the magnitude of the forcing errors (Fig. 6).

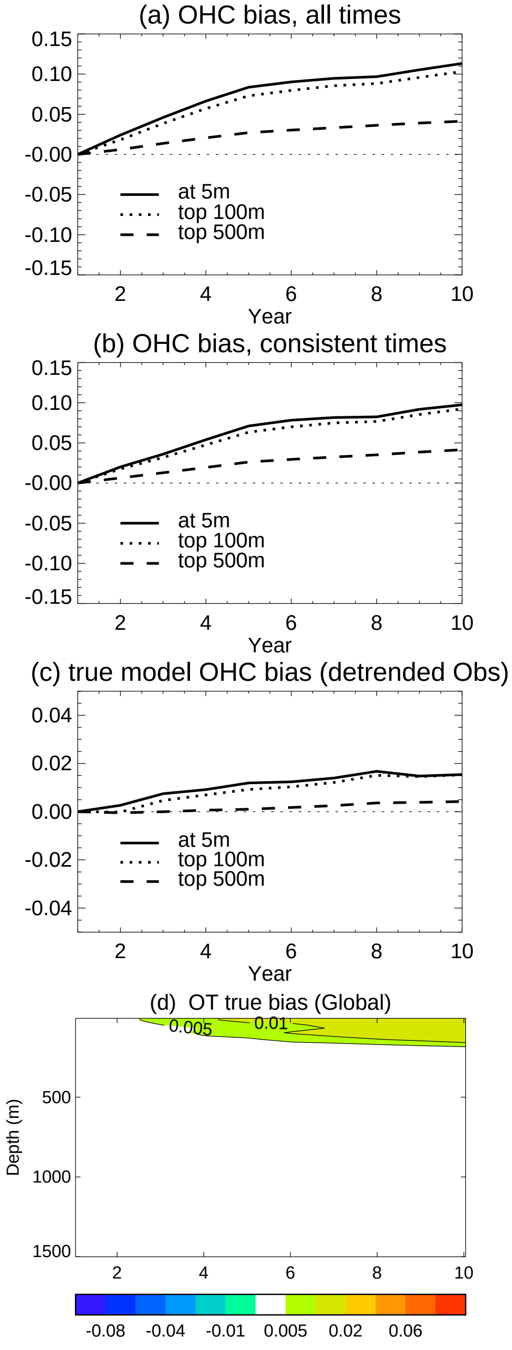

FIG. 11. Time evolutions of ensemble mean bias tendencies (K) for ocean temperature at 5m and ocean heat content (top 100m and top 500m) of HadCM3 NoAssim PPE hindcasts for the period 1961-2010 against Met Office ocean analysis data. (a) using all verification times (1961-2010), (b), using consistent verification times (1971-2001), (c) true bias tendency with linear trend removed in the analysis before calculating bias tendency associated with observed variability. (d) Time evolution of ensemble mean true bias tendency (K) as a function of depth for global ocean temperature for HadCM3 NoAssim PPE hindcasts for the period 1961-2010 against the Met Office ocean analysis.

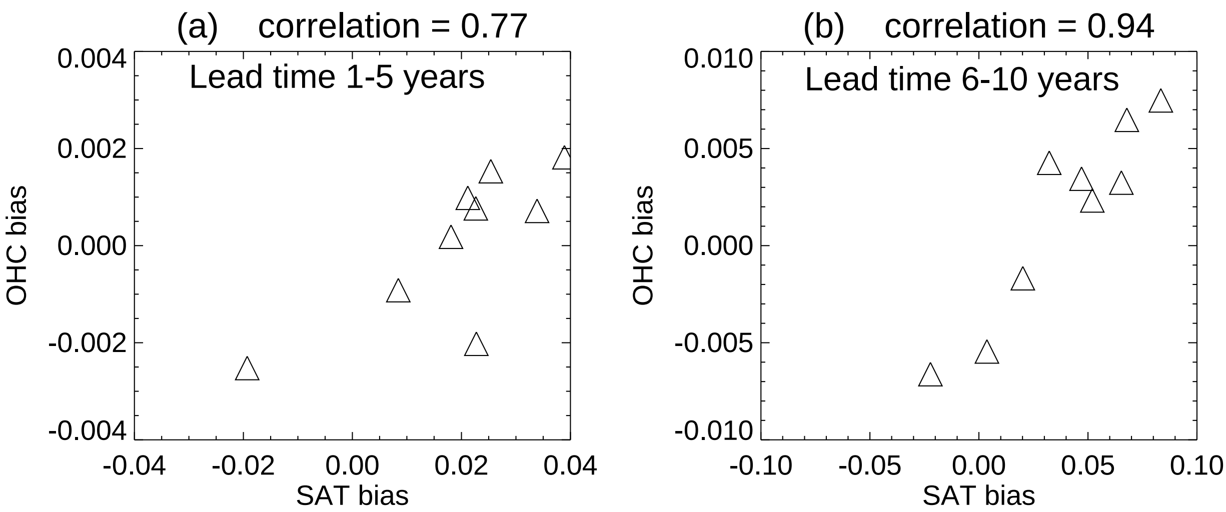

FIG. 12. Relationships between global mean SAT true bias tendencies (K) (against HadCRUT4 data) and global mean OHC (top 1000m) bias tendencies (against the Met Office ocean analysis) for 9 PPE model versions. (a) average for lead years 1-5, and (b) average for lead years 6-10.

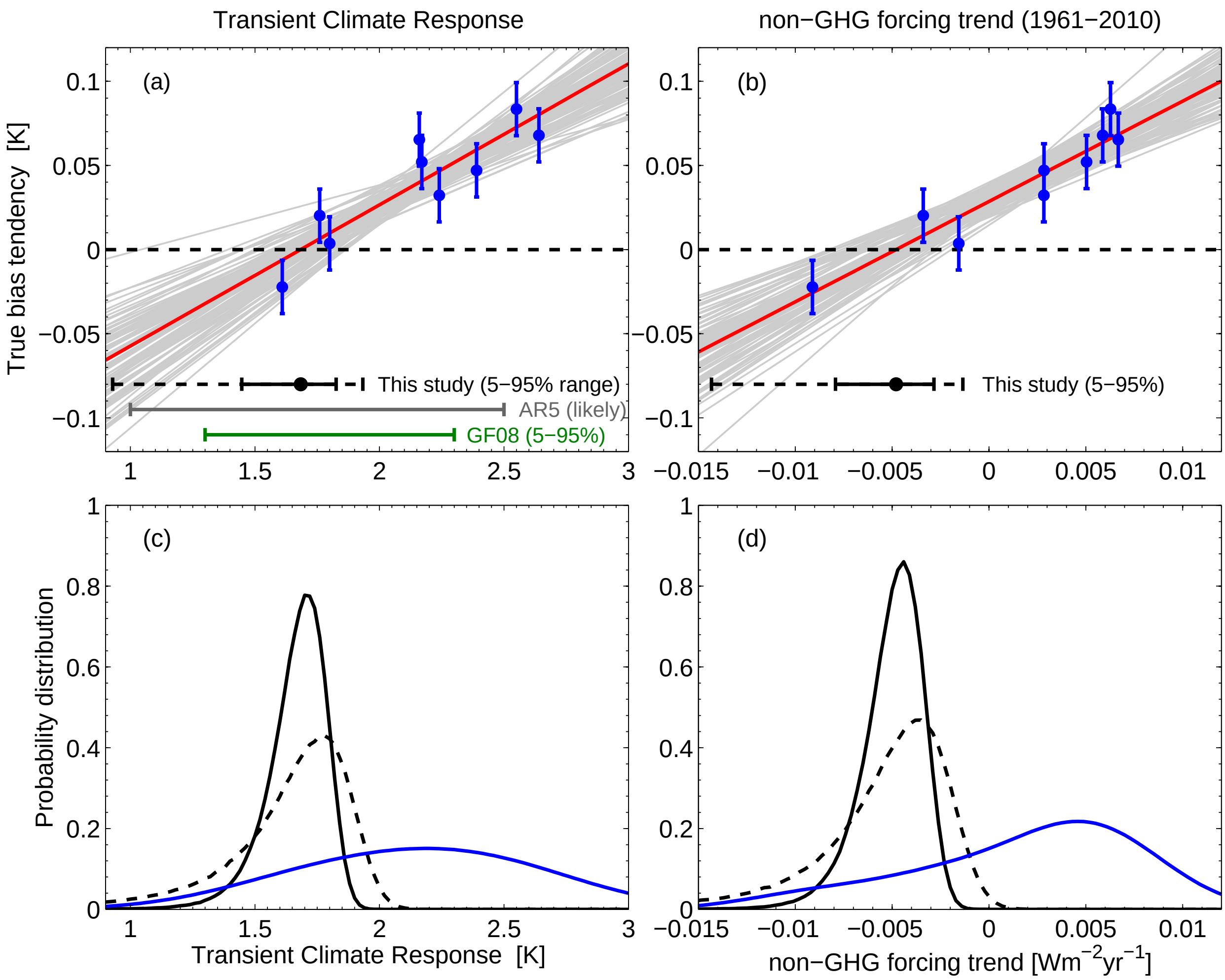

FIG. 13. Relationships between the lead years 6-10 averaged global mean SAT true bias tendencies (K) against HadCRUT4 data for each version of PPE hindcasts for (a) TCR and (b) non-GHG aerosol forcing trend, using 9 PPE model versions. The error bars for bias tendency are based on the toy model (Fig. 7). Grey lines are example linear fits to TCR and to the non-GHG aerosol forcing trend using a Monte-Carlo approach, and the red lines are the best fit. The constrained ranges of TCR and the non-GHG aerosol forcing trend are shown as black bars assuming a true bias tendency error of 0.016K (solid) and 0.032K (dashed). Other ranges for TCR (Stocker et al. 2013,Gregory and Forster 2008 - denoted GF08) ranges are also given. (c,d) estimated probability distribution functions (PDFs) of unconstrained (blue) and constrained (full black and dotted black) TCR and non-GHG aerosol forcing trends. The dashed black lines indicate the PDF for doubled uncertainties in the true bias tendency.