AGU100 ADVANCING EARTH AND SPACE SCIENCE

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001838

Special Section:
Geophysical Fluid Dynamics Laboratory CMIP6 Models

Key Points:
- CM4.0 has high equilibrium and transient climate sensitivities, both near the CMIP5 75th percentile
- The energy budget sensitivity estimation method significantly underestimates CM4.0's sensitivities
- CM4.0's excessive warming over recent decades most likely indicates excessive transient sensitivity

Correspondence to:
M. Winton,
Michael.Winton@noaa.gov

Citation:
Winton, M., Adcroft, A., Dunne, J. P., Held, I. M., Shevliakova, E., Zhao, M., et al. (2020). Climate sensitivity of GFDL's CM4.0. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001838. https://doi.org/10.1029/2019MS001838

Received 25 JUL 2019
Accepted 17 DEC 2019
Accepted article online 17 DEC 2019

Climate Sensitivity of GFDL's CM4.0

M. Winton[1], A. Adcroft[2], J. P. Dunne[1], I. M. Held[2], E. Shevliakova[1], M. Zhao[1], H. Guo[1], W. Hurlin[1], J. Krasting[1], T. Knutson[1], D. Paynter[1], L. G. Silvers[3,4], and R. Zhang[1]

[1]GFDL/NOAA, Princeton, NJ, USA, [2]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA, [3]Cooperative Institute for Modeling the Earth System, Princeton University, Princeton, NJ, USA, [4]School of Marine and Atmospheric Science, State University of New York at Stony Brook, Stony Brook, NY, USA

Abstract GFDL's new CM4.0 climate model has high transient and equilibrium climate sensitivities near the middle of the upper half of CMIP5 models. The CMIP5 models have been criticized for excessive sensitivity based on observations of present-day warming and heat uptake and estimates of radiative forcing. An ensemble of historical simulations with CM4.0 produces warming and heat uptake that are consistent with these observations under forcing that is at the middle of the assessed distribution. Energy budget-based methods for estimating sensitivities based on these quantities underestimate CM4.0's sensitivities when applied to its historical simulations. However, we argue using a simple attribution procedure that CM4.0's warming evolution indicates excessive transient sensitivity to greenhouse gases. This excessive sensitivity is offset prior to recent decades by excessive response to aerosol and land use changes.

Plain Language Summary We evaluate the climate sensitivity of the Geophysical Fluid Dynamics Laboratory (GFDL) CM4.0 climate model. Climate sensitivity is an important factor determining the magnitude of future climate change under anthropogenic forcing. We find that CM4.0 is a high climate sensitivity model. A simple method for estimating climate sensitivity from historical changes significantly underestimates CM4.0's sensitivity when applied to CM4.0's historical simulation. However, more sophisticated methods that make use of the detailed evolution of global warming identify CM4.0 as most likely too sensitive to anthropogenic forcing.

1. Introduction

CM4.0 is the latest in a series of global climate models produced by the Geophysical Fluid Dynamics Laboratory (GFDL), developed in part to participate in a climate model intercomparison project (CMIP). CMIP model results are one of the bases for a series of comprehensive reports produced by the Intergovernmental Panel on Climate Change (IPCC) to interpret past climate changes and project future climate change with a focus on this century. CM4.0 is driven by prescribed greenhouse gas concentrations, including CO_2. Ozone and other atmospheric oxidants are also prescribed, while aerosols are simulated from emissions or emissions of precursors. Land vegetation is interactive, while glaciers and icecaps are prescribed. Improvements of the climatology and variability simulations of GFDL's CMIP6 generation CM4.0 have been documented by Held et al. (2019).

Model simulations are provided to the CMIP archive with the objective of projecting climate changes on the decadal to centennial timescales making it important to assess the model's climate sensitivity, its prediction of climate change under a given forcing. Since future forcing scenarios, as well as historical forcing estimates, are uncertain and evolving, idealized atmospheric CO_2 increase experiments have been used over the history of climate modeling to measure model sensitivity to a benchmark radiative forcing. CO_2 forcing is appropriate for this purpose because the centrality of fossil fuels to industrial economies and the long atmospheric lifetime of CO_2 make it very likely that it will increasingly dominate other radiative forcings. In addition to an 1850 control ("piControl") experiment, the CMIP6 DECK experiments contain two idealized experiments: an abrupt quadrupling of atmospheric CO_2 ("abrupt-4xCO2") and an increase of 1% per year ("1pctCO2"). Since CO_2 radiative forcing is nearly logarithmic in concentration, the 1pctCO2 experiment gives a nearly linear rise in CO_2 radiative forcing. We will use these CMIP6 DECK experiments along with the CMIP6 historical experiment and several custom diagnostic experiments to evaluate and critique CM4.0's climate sensitivity.

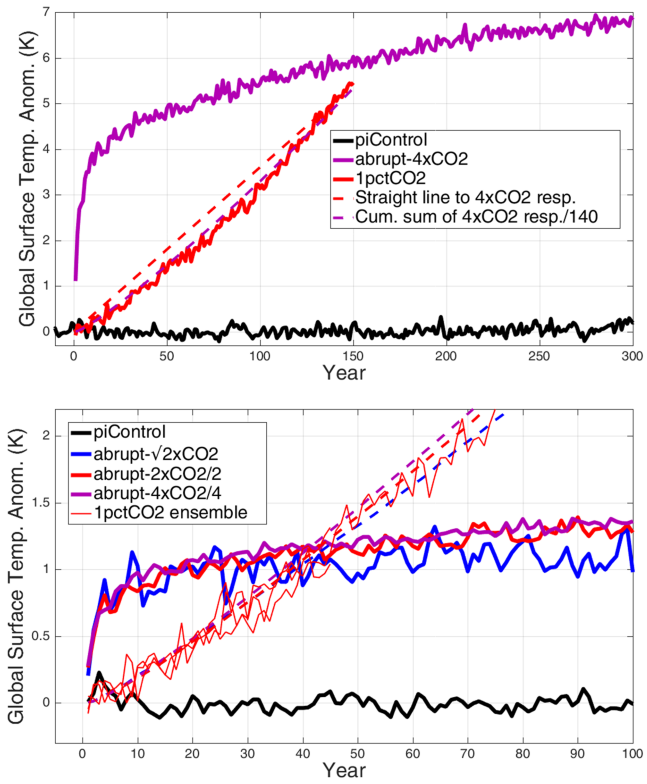

Figure 1. piControl, 1pctCO2, and abrupt CO_2 increase experiments: (a) standard CMIP6 DECK experiments, and (b) additional experiments: abrupt $\sqrt{2}$ times, doubled and quadrupled CO_2 rescaled to $\sqrt{2}$ times CO_2, and a three-member ensemble of 1pctCO2 experiments. Linear estimates of the response to linear forcing using equation (1) are shown as dashed purple line in (a) and dashed purple, red, and blue lines in (b).

The warming at CO_2 doubling in the 1pctCO2 experiment (Year 70) is designated the transient climate response (TCR) and is the primary measure of climate sensitivity under increasing forcing. CO_2 doubling corresponds roughly to the forcing expected near the middle of this century, while quadrupled CO_2 forcing corresponds to the forcing at the end of the century with high-end emissions. Gregory et al. (2015) found that, among a group of CMIP5 models, there was a 0.9 correlation between warming in response to 1pctCO2 forcing at quadrupling and that projected under RCP8.5 forcing at the end of this century. Large ranges of TCR and projected warming—typically at least 50% of the mean—are simulated by the models. This uncertainty propagates from the physical simulation to become uncertainty about impacts and the mitigations needed to avoid them. The regional pattern of temperature change and changes in other variables are also important for impacts, but since transient temperature change under greenhouse gas forcing is generally well approximated as a static pattern function times the global mean surface temperature change time series (Tebaldi & Arblaster, 2014), and because many other climate variables and impacts are related to global mean temperature change, large uncertainty in this quantity effectively prohibits accuracy of regional changes and impact projections. Accuracy of both global temperature and regional patterns—of Arctic amplification, for example—is necessary for accurate regional temperature projections.

The TCR has been assessed by the IPCC since the 2001 report without significant changes in the range. An older sensitivity measure, the equilibrium climate sensitivity (ECS), has a similarly large uncertainty that has also not been reduced over time, in spite of dramatic improvements in model resolution, comprehensiveness, and quality of the simulated climatology and variability. Meanwhile, climate change has progressed to the point where the warming from the preindustrial era, mostly attributed to anthropogenic forcing, is about half the model mean TCR. Given that substantial forced climate change has already occurred, it is reasonable to expect that historical observations will increasingly constrain climate sensitivity. In this study we draw connections between GFDL-CM4.0's climate sensitivity characteristics and its historical simulations, looking to constrain the former with the latter, with a particular focus on the TCR. In the next section we review CM4.0's climate sensitivity characteristics and their connections to its historical simulations. In the third section, we look for constraints on sensitivities from simulated preindustrial to present-day differences as has been done in recent studies using the "energy budget method" (Lewis & Curry, 2015, 2018; Otto et al., 2013). We find instead that energy budget method sensitivities are inaccurate when applied to CM4.0 in a perfect model test. In the fourth section, we employ simplified detection/attribution strategies (Gillett et al., 2012) to argue that CM4.0 is most likely not consistent with the historical record and to constrain the TCR using the shape of the historical warming evolution. Analysis of the role of individual feedbacks in this inconsistency is left to future work. We summarize the results in the final section.

2. CM4.0's Sensitivity Characteristics

The global mean surface air temperature changes for CM4.0's idealized DECK experiments are shown in Figure 1a. The abrupt-4xCO2 experiment has been extended from the 150 years requested by the CMIP6 protocol to 300 years in order to make a more accurate assessment of the ECS. The TCR from the difference of 20-year averages of the 1pctCO2 and piControl experiments centered on Year 70 is 2.05 (±0.10) K. This value would place CM4.0 near the 75th percentile of the CMIP5 model TCRs. Since we will be comparing the TCR to observational estimates that use a combination of sea surface (SST) and air temperature to estimate global surface temperature (Richardson et al., 2016), we also calculated the TCR using SST over the ice-free ocean and surface air temperature elsewhere. We found that CM4.0's TCR was reduced by less than 0.1 K using this procedure. For simplicity and comparability with previous work we ignore this slight discrepancy and use global surface air temperature as our standard measure.

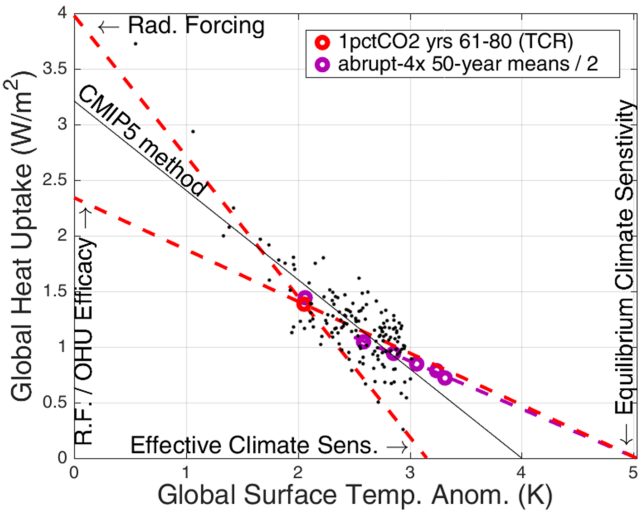

Figure 2. CM4.0 Gregory diagram showing the climate state (global surface temperature/heat uptake change) at 1pctCO2 doubled CO_2 (red) and abrupt-4xCO2 50-year averages (purple, divided by 2). Annual values (divided by 2) for the first 150 years of the abrupt CO_2 quadrupling experiment are shown (black dots) along with the regression line as used by Andrews et al. (2012) to calculate the CMIP5 ECSs. Dashed red lines from the radiative forcing and ECS through the doubled CO_2 state define the Effective Climate Sensitivity and Ocean Heat Uptake Efficacy, respectively.

CM4.0's warming at CO_2 quadrupling (Year 140 in the 1pctCO2 experiment) is 5.0 K, well above twice its TCR. Gregory et al. (2015) found a similar nonlinearity in a group of CMIP5 models: The warming from doubling to quadrupling was 40% larger than the TCR in the multimodel mean, a little less than we find here for CM4.0. Using the "Hansen method" (Chung & Soden, 2015; Hansen et al., 2005) with the atmospheric component of CM4.0, AM4.0 (Zhao et al., 2018a, 2018b), we find that the radiative forcing of the first CO_2 doubling from preindustrial is 3.98 W/m^2 while that of the second is 3.93 W/m^2. Therefore, the extra warming is due to the response, in particular ocean effects that we will discuss later.

Curvature in the response to a linear ramp in CO_2 forcing is not necessarily a signature of nonlinearity of the system, since it can be emulated with a linear response function. In fact, Gregory et al. (2015) found that this curvature in the response to ramp CO_2 forcing in the CMIP5 models could mostly be accounted for using a linear response calculated from the abrupt-4xCO2 experiment. The calculated linear response in that study had slightly less curvature than the full response leaving room for a modest contribution from nonlinear response to forcing. It can be shown that the response to linearly increasing forcing, assuming a linear response to forcing, is proportional to the integral of the step response:

$$T_{\text{linear}}(t) = \frac{1}{t_{\text{step}}} \int_0^t T_{\text{step}}(\tau)\mathrm{d}\tau \quad (1)$$

where t_{step} is the time required for the linearly increasing forcing to achieve the magnitude of the step forcing (140 years, in this case). This integral is also plotted in Figure 1a. The comparison shows that the CM4.0's warming is slightly more curved than expected for a linear system, similar to the result for the CMIP5 model mean.

For interpretation of CM4.0's historical simulations we are interested in the early part of the 1pctCO2 experiment where the warming is less than 1 K and the forcing less than about $\sqrt{2}$ times preindustrial CO_2, halfway to doubled CO_2 forcing on the log scale. To help estimate the forced response during this period, we have performed two additional 1pctCO2 ensemble members integrated to Year 45. These were initialized at Years 40 and 82 of the control simulation. Motivated by the evidence of nonlinearity in the 1pctCO2 experiment, we have also performed two additional abrupt CO_2 increase experiments with double and $\sqrt{2}$ times preindustrial CO_2. Rescaling all of the abrupt CO_2 increase experiments to the $\sqrt{2}$ times forcing shows that the responses are indeed slightly nonlinear in forcing level with the largest effect occurring between $\sqrt{2}$ times and doubled CO_2 levels (Figure 1b). The integrals of the step responses show that nonlinear effects might account for warming of about 0.1 K between $\sqrt{2}$ times CO_2 and doubling.

Figure 2 shows a plot of global temperature change against top-of-atmosphere heat uptake for CO_2 doubling in the 1pctCO2 experiment and rescaled 50-year means for the abrupt-4xCO2 experiment. This "Gregory diagram" (Gregory et al., 2004) is very useful for estimating the ECS and other climate sensitivity parameters. We estimate the ECS by extrapolating the abrupt-4xCO2 climate states to zero heat uptake, ignoring the first 50 years due to contamination by a fast time scale adjustment (Held et al., 2010). This method yields an ECS estimate of 5.0 K. The method was used by Winton et al. (2013) to obtain ECS estimates for GFDL-ESM2M and CM3 that were within 0.2 K of the values obtained by integrating these models for several thousand years to their actual equilibria (Paynter et al., 2018). The extrapolation assumes that equilibration is effected by processes that are already active over the first 300 years. Li et al. (2013) show a case where this assumption is violated as the temperature/heat uptake relationship changes slope after 1,200 years of integration leading their extrapolation to overestimate ECHAM5's equilibrium response to doubling by 0.7 K.

A different extrapolation method was used to obtain the model ECS values cited in the IPCC fifth report (Andrews et al., 2012). This extrapolation used the first 150 years of annual values from the abrupt-4xCO2 experiment. The method underestimated the true ECS values for GFDL-ESM2M and CM3 by 0.9 and 0.8 K, respectively (Paynter et al., 2018). Applying the Andrews et al. method to CM4.0, we find that it gives an

ECS estimate of 4.0 K, 1 K lower than our extrapolation method ECS of 5.0 K. Comparing CM4.0's Andrews et al. method ECS to those compiled in their paper shows that CM4.0 is at about the 75th percentile, similar to its TCR placement in that group. By these conventional sensitivity measures CM4.0 is in the middle of the more sensitive half of the models.

The effective climate sensitivity, ECS_{2X}, approximates the ECS by scaling up the transient warming with the ratio of the doubled CO_2 radiative forcing to transient forcing minus heat uptake (Murphy, 1995):

$$ECS_{2X} = \Delta T \frac{R_{2X}}{R - N} \tag{2}$$

where ΔT is the global temperature change, R is the radiative forcing, R_{2X} is the doubled CO_2 radiative forcing, and N is the earth energy imbalance or heat uptake. The scaling up can be justified with the linear perturbation energy balance:

$$R = \lambda \Delta T + N \tag{3}$$

noting that $N = 0$ at equilibrium. In equation (2), the radiative damping parameter λ is the negative of the radiative feedback (positive for negative feedback). The effective sensitivity based on the transient state at CO_2 doubling in the 1pctCO2 experiment ($\Delta T = TCR$ and $R = R_{2X}$) can be visualized in Figure 2 as the intersection of the line from the radiative forcing through the doubled CO_2 climate state with the x axis. CM4.0's effective sensitivity of 3.2 K likely underestimates its ECS substantially. The underpinning assumption, from equation (3), that the radiative forcing and heat uptake have equivalent impacts on temperature, does not apply to CM4.0, leading to a large low bias. If heat uptake has greater impact than CO_2 forcing the effective sensitivity would be expected to increase over time as the climate equilibrates with its forcing. This effect, first noted by Senior and Mitchell (2000), was found by Armour (2017) to be generally true of the CMIP5 models in CO_2 increase experiments.

We can read two important parameters for the ocean's role in transient climate change from the Gregory diagram in Figure 2. The ocean heat uptake *efficiency*

$$\gamma = \frac{N}{\Delta T} \tag{4}$$

is 0.67 $W \cdot m^{-2} \cdot K^{-1}$ in 1pctCO2 at CO_2 doubling, near the CMIP5 model mean of 0.64 $W \cdot m^{-2} \cdot K^{-1}$ (Kuhlbrodt & Gregory, 2012). The second parameter, the ocean heat uptake *efficacy*, is the ratio of global temperature responses to ocean heat uptake and to CO_2 forcing changes (Winton et al., 2010). The ocean heat uptake efficacy can be written in terms of the TCR and ECS as

$$\epsilon = \frac{(ECS - TCR)/N}{ECS/R_{2X}} \tag{5}$$

based on a modified form of equation (3) that allows for different impacts of radiative forcing and ocean heat uptake on temperature,

$$R = \lambda \Delta T + \epsilon N. \tag{6}$$

Several recent studies have noted that climate feedback varies with the SST response pattern (Andrews et al., 2015; Armour et al., 2013; Gregory & Andrews, 2016; Zhou et al., 2017). Ocean heat uptake efficacy parsimoniously captures this effect by constructing the feedback from two response components having different SST patterns—one associated with the equilibrium response to the forcing and the other with the ocean heat uptake. Equation (5), for example, can be rewritten in this form as

$$\lambda_{2X} = \lambda_{EQ} + (\epsilon - 1)\gamma_{2X} \tag{7}$$

where $\lambda_{2X} = R_{2X}/TCR - \gamma_{2X}$, $\lambda_{EQ} = R_{2X}/ECS$, and $\gamma_{2X} = N/TCR$ is the heat uptake efficiency at CO_2 doubling. To the extent that ϵ is relatively constant, as it is in CM4.0's abrupt-4xCO2 experiment after the first few decades, equation (7) will pertain to other times besides CO_2 doubling, thereby capturing the time variation of the feedback with a constant parameter.

Most models have ocean heat uptake efficacy values greater than 1 indicating that ocean heat uptake forces more temperature change than CO_2 per W/m^2 and the transient climate feedback is larger (more damping)

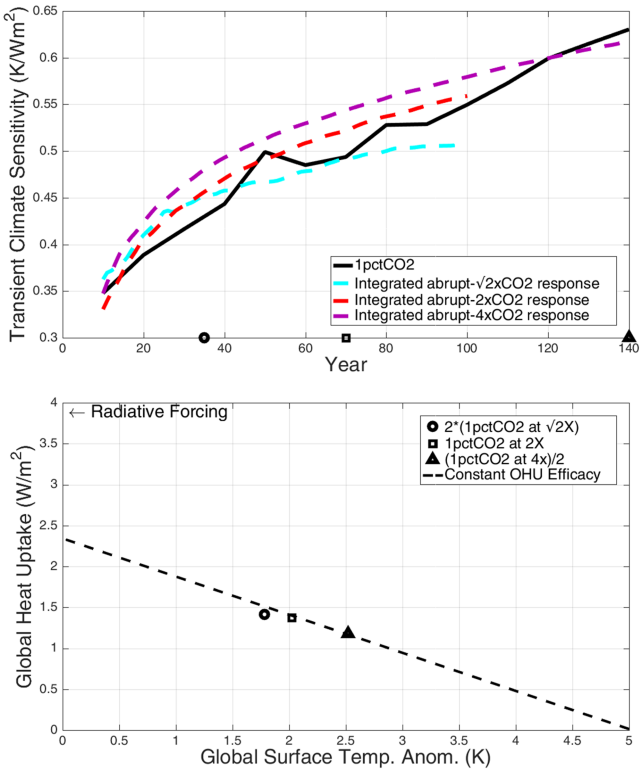

Figure 3. (a) Time variation of transient climate sensitivity (solid) with linear estimates using equation (1) (dashed) and (b) Gregory diagram showing 1pctCO2 transient states for forcing levels corresponding to Years 35, 70, and 140 normalized to doubled CO_2 for comparison.

than the equilibrium feedback. In Figure 2 the efficacy value is the ratio of the radiative forcing to the *y* intercept of the line from the ECS through the doubled CO_2 climate state, 1.7 for CM4.0. This value is considerably larger than the multimodel mean value of 1.3 found by Winton et al. (2010). CM4.0's large ocean heat uptake efficacy is a major factor in the underestimation of its ECS by the CMIP5 method and by the effective sensitivity. ECS estimates for models with lower ocean heat uptake efficacy should be less sensitive to method.

Figure 3a shows the transient climate sensitivity (TCS), the warming divided by the forcing, along the course of the 1pctCO2 experiment. The linear model expectation for TCS from integrating abrupt CO_2 increase responses is also shown. The TCS rises significantly over the course of the experiment. It rises across the values expected for linear responses at particular forcing levels (the dashed curves) which themselves rise significantly with time. At Year 35 the atmospheric CO_2 level is $\sqrt{2}$ times preindustrial, near the present-day level. From this level, the TCS rises by 15% to CO_2 doubling (Year 70) and 47% to quadrupling (Year 140). This suggests that estimates of future warming based on the present-day TCS, even if this can be accurately estimated, may be low biased, especially for century-scale projections. Nonlinearity of the response—greater sensitivity at higher levels of forcing—contributes the part of this behavior evident in the crossing of the response curves. The rising of the response curves themselves, due to the warming impact of reducing ocean heat uptake relative to radiative forcing, contributes the other part of the sensitivity increase.

We can examine the increasing TCS over 1pctCO2 due to heat uptake equilibration using a scaled Gregory diagram (Figure 3b). Here we compare the climate states for $\sqrt{2}$ times, doubled, and quadrupled CO_2 levels in the 1pctCO2 experiment. The states have all been normalized, based on their log-CO_2 forcing levels, to doubled CO_2. Without equilibration differences between the states, they would all lie on top of each other on this plot. Instead they track a line of constant ocean heat uptake efficacy as time progresses and the forcing increases, similarly to the equilibration pathway under constant quadrupled CO_2 forcing (Figure 2). This diagram shows that some of the extra warming over the second CO_2 doubling relative to the first in the 1pctCO2 experiment is due to reduced (scaled) heat uptake moving the climate state toward its (scaled) equilibrium. Because this equilibration takes place along a line of constant ocean heat uptake efficacy rather than one of constant effective sensitivity (emanating from the radiative forcing on the *y* axis), reductions in both ocean heat uptake efficiency and radiative damping contribute to this extra warming.

The heat uptake efficiency reduction during equilibration is associated with ocean warming below the mixed layer. We expect that this warm anomaly would persist even if the forcing were returned to preindustrial. For this reason, Held et al. (2010) referred to this component of the surface warming as "recalcitrant warming." Recalcitrant warming increases over time as the climate equilibrates with its forcing and it is a part of the TCR that may be underestimated when scaling up the present-day warming to produce observation-based TCR estimates.

We can demonstrate that there is extra recalcitrant warming at CO_2 doubling relative to present day in CM4.0 directly using return-to-preindustrial-experiments following Held et al. (2010). The intent of these experiments is to expose the component of the warming that is due to sub-mixed-layer ocean warming by abruptly removing the component due to current radiative forcing. The return to preindustrial forcing experiments from 2015 in the historical experiment and Year 70 of the 1pctCO2 experiment are shown in Figure 4. As expected, there is a small difference of a few tenths of kelvins in the warming that remains after radiative forcing is removed. This difference is warming at CO_2 doubling that will be missed if the TCR is estimated by scaling up the present-day warming with radiative forcing. CM4.0's present-day warming does not include any recalcitrant warming to scale up.

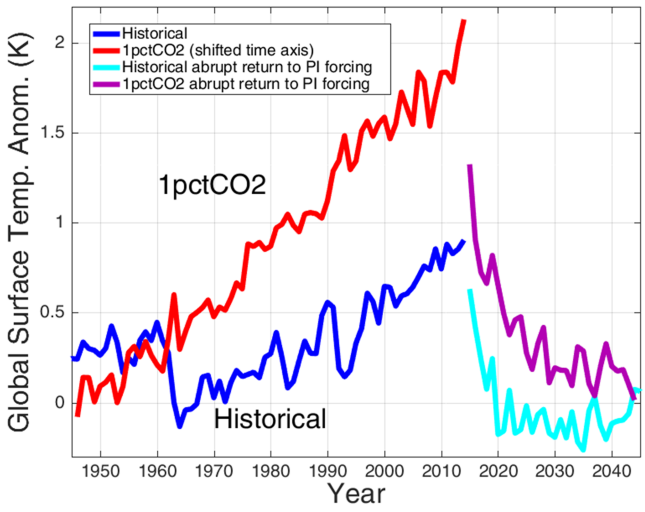

Figure 4. Global surface temperature in historical (blue) and 1pctCO2 (red) increase experiments and their corresponding return-to-piControl-forcing experiments (light blue and purple, respectively).

We will show below that the historical CO_2-only forced response can be used to quantify the present-day recalcitrant warming and other climate change characteristics. A direct way to calculate this response would be to make a historical CO_2-only experiment. We have chosen instead to perform an alternative control for the historical experiment using all historical forcings except CO_2, which is fixed at its 1850 value. The difference of the all-forced historical and this experiment is taken to be the CO_2-only response. The reason for this approach is to allow carbon fertilization to be calculated as a difference between land carbon content in the two experiments both having historical land use change, although we do not make use of this feature of the experiment in the current study. More generally, to the extent that there are interactions between responses to individual forcing agents, the all-forced historical makes a better control experiment. Below we use "CO_2-only" to refer to differences between the all-forced historical ensemble and the fixed-CO_2 historical.

Figure 5 shows the all-forced and fixed-CO_2 historical experiments. We have performed three historical ensemble runs branched from Years 250, 290, and 332 of the preindustrial control. In general, the CO_2 forced warming can be either less than or greater than the total depending upon the balance of other forcings. Since the fixed-CO_2 historical temperature perturbation is nearly 0 after 2000, the responses to other forcings cancel out and nearly all of CM4.0's warming can be thought of as due to CO_2. A larger CO_2-forced component of present-day warming should result in greater future warming for two reasons: It indicates a larger climate sensitivity, and secondly, a larger fraction of the current warming will be long-lived because it is associated with this long-lived forcing agent. The historical CO_2 TCS, based on warming and forcing averaged over 1995–2014, is 0.45 $KW^{-1}m^2$ similar to the values early in the 1pctCO2 experiment (Figure 3a) and 14% less than the value at doubled CO_2. This confirms our expectation from the early part of the 1pctCO2 experiment that TCR estimates based on present-day observations will be slightly low biased. A more general assessment of this relationship should become possible with the CMIP6 DAMIP CO_2-only historical experiment.

In summary, CM4.0 has above average transient and equilibrium climate sensitivities (TCR = 2.05 K, ECS = 5.0 K), the former in spite of having an above average ocean cooling effect through its large ocean heat uptake efficacy. CM4.0's sensitivity appears to increase over time as the radiative forcing increases due to declining ocean heat uptake efficiency in combination with its large efficacy, and a more sensitive response to the larger forcing levels.

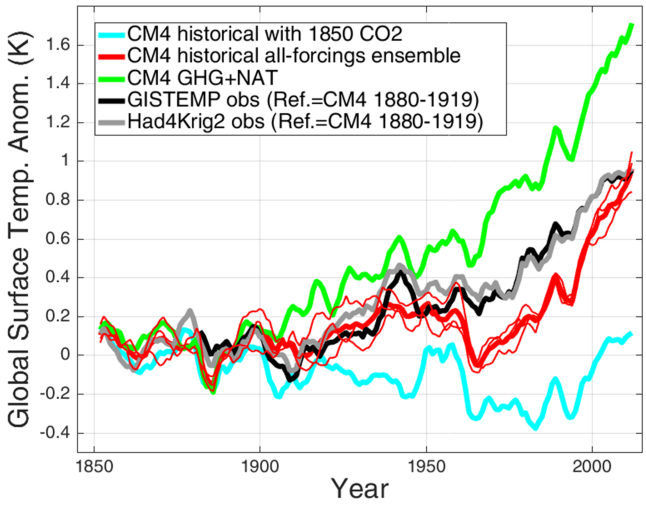

Figure 5. Historical global surface temperature observations (black and gray) and CM4.0 simulations with historical forcing (ensemble members in red; mean is thickened), historical except fixed preindustrial CO_2 (light blue), and historical except fixed preindustrial aerosol emissions and land use (GHG+NAT, green). All series are 5-year smoothed.

3. Constraints From Present-Day Changes

Figure 5 shows CM4.0's response to historical greenhouse gas and natural forcing. These forcings induce nearly twice the warming of the total historical forcing. The major cooling agents, aerosols and land use, account for the difference. It is the potentially compensating influences of the uncertain climate sensitivity and magnitudes of the cooling forcings that prevents the climate sensitivity from being accurately determined from the historical warming (Kiehl, 2007).

Recently, several studies have estimated TCR and ECS using present-day observations of temperature change and heat uptake and assessments of radiative forcing (Lewis & Curry, 2015, 2018; Otto et al., 2013). These studies use an energy budget method that scales up present-day temperature change to transient or equilibrium CO_2 doubling using the present-day radiative forcing, the doubled CO_2 forcing, and the heat uptake as follows:

$$TCR_{EB} = \Delta T_{ALL} \frac{R_{2X}}{R_{ALL}} \quad (8)$$

Table 1
Energy Budget Method Parameters Using Observations (Lewis & Curry, 2018), CM4.0 Historical Ensemble Means, and the CM4.0 Historical CO_2-Only Response

	Lewis and Curry (2018)	CM4.0 all-forced	CM4.0 CO_2-forced
ΔT (K)	0.79 (0.63–0.94)	0.65	0.71
R (W/m^2)	2.26 (1.44–3.09)	2.04	1.58
TCR_{EB} (K)	**1.32** (0.95–2.0)	**1.27**	1.78
$\frac{TCR_{EB}}{TCR}$	—	0.62	0.87
N (W/m^2)	0.49 (0.29–0.69)	0.56	0.55
ECS_{EB} (K)	**1.69** (1.35–2.25)	**1.76**	2.73
$\frac{ECS_{EB}}{ECS}$	—	0.35	0.54

Note. Numbers in bold are observation-based and model TCR and ECS estimates. The listed Lewis and Curry (2018) results refer to their differences between 1995–2016 and 1869–1882 averages with temperatures taken from the Had4KrigV2 data set. CM4.0's actual TCR and ECS are 2.05 and 5.0 K, respectively.

$$ECS_{EB} = \Delta T_{ALL} \frac{R_{2X}}{R_{ALL} - N_{ALL}} \tag{9}$$

The "ALL" subscript refers to quantities evaluated at present day when all forcing agents are active. For equation (8) to hold generally, the heat uptake efficiency and efficacy must be the same in the historical and TCR states as can be seen by combining equations (4) and (6) to obtain

$$\Delta T = \frac{R}{\lambda + \epsilon\gamma}. \tag{10}$$

If this is the case, equation (10) allows transient temperature changes to scale with radiative forcing. Equation (9) is simply the effective climate sensitivity (equation (2)) obtained by scaling up the historical warming.

Observation-based energy budget method sensitivity estimates are generally skewed low compared to CMIP model TCR and ECS distributions. For example, the Lewis and Curry (2018) most likely values of 1.2 K and 1.5 K for the TCR and ECS, respectively, are much lower than their counterpart CMIP5 model averages of 1.8 and 3.2 K. GFDL-CM4.0, as a more sensitive model, has an even larger discrepancy with the energy budget method estimates. In this section, we use CM4.0's historical simulation of ΔT, R, and N to generate TCR_{EB} and ECS_{EB} and compare them to the observationally derived estimates and the model's actual values. The former comparison tests the fidelity of CM4.0's historical simulation and latter tests the energy budget method itself.

Table 1 shows the Lewis and Curry (2018) values using the globally interpolated Had4KrigV2 data set (Cowtan & Way, 2014). This data set is used for direct comparison to CM4.0's global mean temperature changes. From the several Lewis and Curry (2018) estimates we use those based on the 1869–1882 preindustrial and 1995–2016 present-day periods. These offer the best comparison to CM4.0's historical simulation because the CMIP6 historical period ends in 2014 and it is desirable to use 20-year, or longer, averaging periods to estimate climate changes. The TCR and ECS estimates using these choices are a little larger than the Lewis and Curry (2018) preferred estimates but still much lower than CM4.0's values. Table 1 also lists values for the CM4.0 all-forced ensemble mean changes and, to remove forcing heterogeneity as a source of sensitivity discrepancy, CM4.0's historical CO_2-forced changes.

CM4.0's historical simulation of the variables used for energy balance method sensitivities—warming, heat uptake, and radiative forcing—are consistent with the Lewis and Curry (2018) observations. Consequently, CM4.0's sensitivities estimated using the energy budget method agree with the Lewis and Curry (2018) estimates to within 0.1 K, while those using the CO_2-only historical values are 0.5 and 1.0 K larger for TCR and ECS, respectively. Using the CO_2-only numbers reduces but does not eliminate the underestimate of the model's true values indicating that only part of the underestimate is due to forcing effects. First, we address the forcing effect portion of the underestimate, quantified as the difference in all- and CO_2-forced responses.

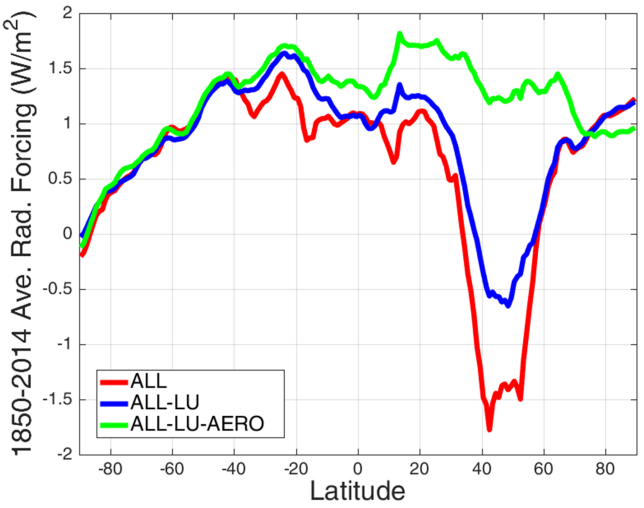

Figure 6. Historical period average forcings calculated for all-forcing (red), all except land use (blue), and all except land use plus aerosol (greenhouse plus natural, green) using AM4.0 top-of-atmosphere radiative flux differences between perturbation and control experiments with the same SST and sea ice boundary conditions.

This discrepancy arises from the efficacy of non-CO_2 forcing agents—that is, their differing responses per W/m^2 forcing. Hansen et al. (2005) showed that the equivalence of a forcing to CO_2 breaks down for forcings that are not well distributed globally. This can be corrected by applying an efficacy factor to a forcing. For example, we can apply an efficacy factor of

$$\epsilon_{ALL} = \frac{\Delta T_{ALL}}{R_{ALL}} \cdot \frac{R_{\mathrm{CO2}}}{\Delta T_{\mathrm{CO2}}} \tag{11}$$

to the present-day radiative forcing, R_{ALL}, to account for its differential impact on warming relative to historical CO_2 forcing, R_{CO2}. Here the "CO2" subscript refers to quantities evaluated at present day under historical CO_2-only forcing.

Figure 6 shows zonal and historical period mean forcings calculated as the top-of-atmosphere net flux in response to the forcing change using AM4.0 with fixed SSTs and sea ice cover (Hansen et al., 2005). The historical greenhouse gas plus natural forcing is fairly evenly distributed meridionally, but land use and aerosol forcings are concentrated on a band between 40°N and 60°N. Since land use and aerosols are similarly distributed we separate the forcings into two clusters: greenhouse gas plus natural (GHG+NAT) and aerosol plus land use (AERO+LU). An experiment with fixed 1850 aerosol precursor emissions and land use was performed to calculate the response to these two forcing clusters, with the AERO+LU response calculated as the difference between all-forcing (ALL) ensemble and GHG+NAT responses.

Figure 7 shows the 1995–2014 global average temperature changes plotted against forcings for the ALL, CO_2-forced and GHG+NAT-forced experiments. The slopes of the lines on the plot, the warming divided by the forcing, are the transient sensitivities. The transient sensitivity for ALL forcing is significantly less than for CO_2—the ratio of the slopes, the all-forcing efficacy, is 0.71. The transient sensitivity for GHG+NAT is nearly the same as that of CO_2-only, giving an efficacy near 1. Under the assumption that the forcings add linearly, the response to AERO+LU forcing is the difference between the ALL and GHG+NAT-forced responses. The AERO+LU transient sensitivity, calculated as this difference divided by the difference in forcing, is larger than for CO_2 forcing, by a factor of about 1.5. A larger-than-unity efficacy is expected for forcings localized at high latitudes and Northern Hemisphere midlatitudes (Hansen et al., 2005; Shindell, 2014). Rotstayn et al. (2015) found that the CMIP5 models had an average aerosol forcing efficacy of 1.4 relative to historical greenhouse gas forcing. Land use forcing efficacies have been estimated both above (Hansen et al., 2005) and below (Davin et al., 2007; Jones et al., 2013) one. Therefore, CM4.0 appears to have an above average aerosol plus land use forcing efficacy leading to a relatively low all-forcing efficacy.

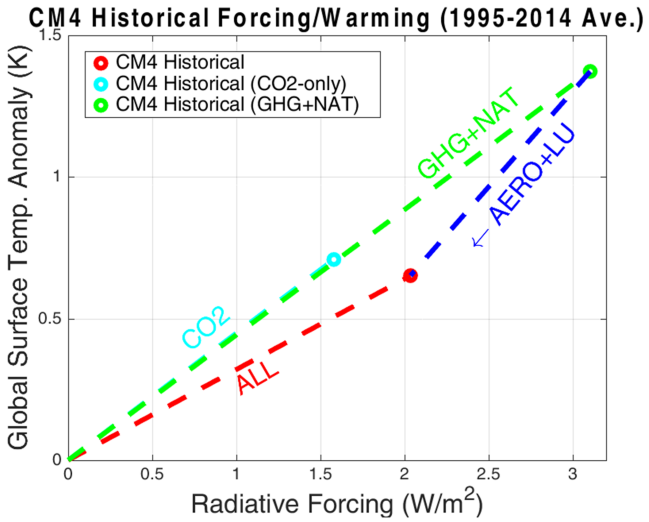

Figure 7. Historical ALL, CO2-only, and GHG+NAT-forced 1995–2014 average radiative forcings and temperature responses. The AERO+LU values are differences between the ALL and GHG+NAT experiments.

Making use of the all-forcing efficacy, we can attribute part of the energy budget method error to the lack of a forcing efficacy factor and the rest to the transient sensitivity increasing effect of ocean warming by using the forcing efficacy adjusted TCR estimate:

$$TCR_{ALL} = \Delta T_{ALL} \frac{R_{2X}}{\epsilon_{ALL} R_{ALL}} \tag{12}$$

After substituting equation (11) we find that

$$TCR_{ALL} = \Delta T_{\mathrm{CO2}} \frac{R_{2X}}{R_{\mathrm{CO2}}} = TCR_{\mathrm{CO2}} \tag{13}$$

Therefore, the TCR estimate based on historical CO_2 forcing divides the error into a part due to forcing efficacy, $TCR_{\mathrm{CO2}} - TCR_{EB}$, and a part due to recalcitrant warming, $TCR - TCR_{\mathrm{CO2}}$. This partition is depicted in

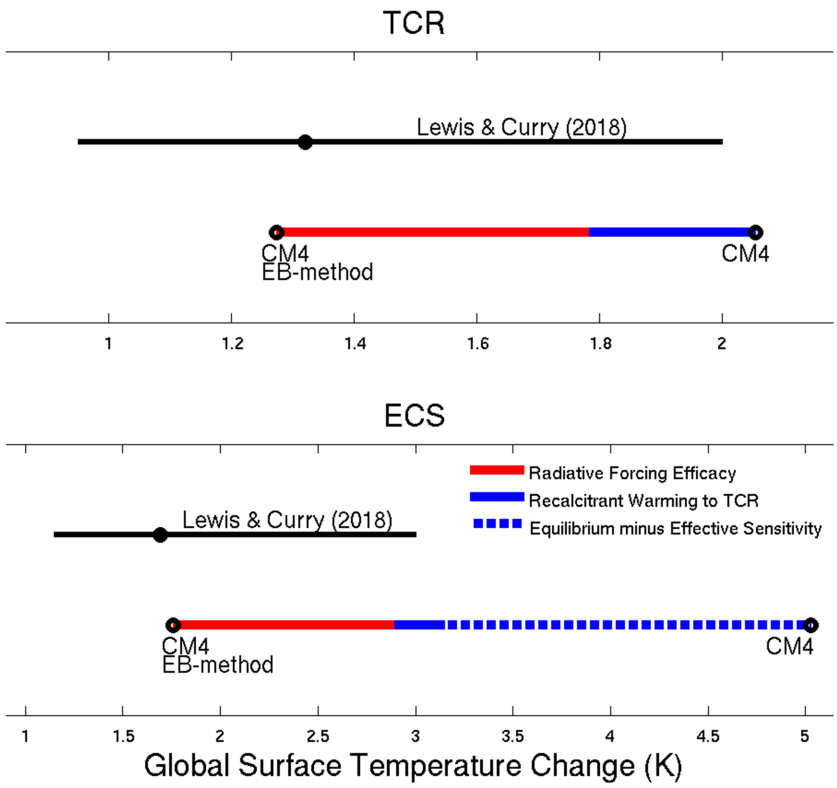

Figure 8. Sources of energy budget method underestimation of CM4.0's (a) TCR and (b) ECS. The Lewis and Curry (2018) energy budget method estimates and 90% confidence intervals are also shown.

Figure 8a. We note that rewriting the residual error after adjusting for radiative forcing efficacy as

$$TCR - \Delta T_{ALL} \frac{R_{2X}}{\epsilon_{ALL} R_{ALL}} = R_{2X} \left(\frac{TCR}{R_{2X}} - \frac{\Delta T_{ALL}}{\epsilon_{ALL} R_{ALL}} \right) \tag{14}$$

clearly shows its association with increased (efficacy adjusted) transient sensitivity at CO_2 doubling relative to present day (the parenthesized term at right).

For equilibrium sensitivity we can quantitatively attribute the energy budget method underestimation of CM4.0's ECS to three factors: forcing efficacy, recalcitrant warming, and ocean heat uptake efficacy, a contributor to the recalcitrant warming (Figure 8b). The first step is to account for historical forcing efficacy similarly to our procedure for TCR:

$$ECS_{ALL} = \Delta T_{ALL} \frac{R_{2X}}{\epsilon_{ALL} R_{ALL} - N_{ALL}} \tag{15}$$

The energy budget method error, $ECS - ECS_{EB}$, can now be partitioned, using ECS_{ALL} and the effective sensitivity ECS_{2X} (equation (2)), as

- $ECS_{ALL} - ECS_{EB}$ due to lack of accounting for forcing efficacy,
- $ECS_{2X} - ECS_{ALL}$ due to extra recalcitrant warming due to ocean warming at transient CO_2 doubling relative to present day, and
- $ECS - ECS_{2X}$ recalcitrant warming entirely due to lack of accounting for ocean heat uptake efficacy larger than 1.

These terms are depicted in Figure 8b. We note that the fact that the climate states follow a line of constant ocean heat uptake efficacy rather than effective climate sensitivity (Figure 3b) as the forcing level increases indicates that some part of the recalcitrant warming to transient CO_2 doubling could also be attributed to ocean heat uptake efficacy. Because this part is already small, we do not partition it further.

To summarize, we have found that the energy budget method gives inaccurate estimates of TCR and ECS for CM4.0 and so does not provide a useful constraint on CM4.0's sensitivities. The lack of forcing efficacy is

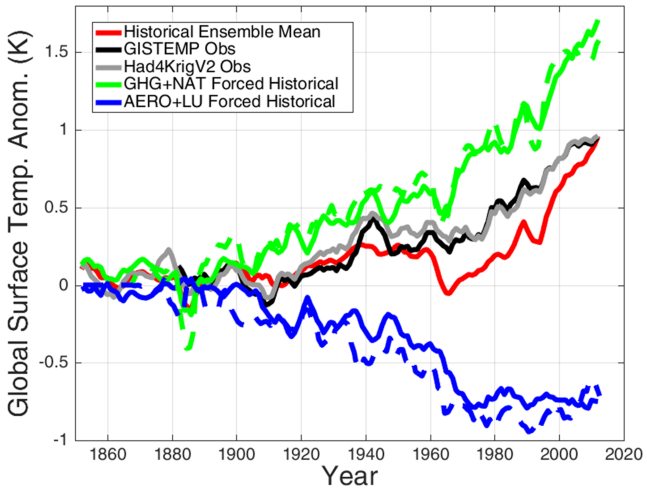

Figure 9. Global temperature response to greenhouse gas plus natural (green) and to aerosol plus land use forcing (blue). Estimates of these responses obtained by applying the transient sensitivity (and aerosol plus land use efficacy) to the radiative forcings for these two components are also shown (dashed). The ensemble mean all-forcing historical warming (red) and observed warming (black and gray) are reproduced here for reference. Five-year smoothing is applied to all series.

the dominant source of energy budget method TCR underestimation and is also important for ECS underestimation. Ocean heat uptake efficacy is the dominant source of ECS underestimation. Thus, we find for CM4.0 that the energy budget method is an unreliable method of estimating sensitivities as has been suggested, more generally, in several recent studies (Armour, 2017; Marvel et al., 2015; Proistosescu & Huybers, 2017). Our finding that ECS_{EB} is lower than the ECS due to radiative and ocean heat uptake efficacies also offers an explanation for the finding of Andrews et al. (2018) that their historical effective sensitivity, $R_{2X}/\lambda_{\text{hist}}$ underestimates the climate model ECSs ($\lambda_{\text{hist}} = -N/\Delta T$ when forcing a model's atmospheric component with historical SST and sea ice changes). The historical SST change pattern incorporates the influences of the efficacies which serve to reduce extratropical warming and hence the increase in radiative damping. Neither radiative forcing efficacy nor ocean heat uptake efficacy affect the radiative damping at a CO_2-forced equilibrium. Consequently, the pattern effect presents an impediment to constraining the ECS with present-day observations (Stevens et al., 2016).

4. A Constraint From Preindustrial to Present-Day Evolution

The last section showed the difficulty of constraining sensitivity using only preindustrial to present-day changes. In this section, we try to constrain TCR using the evolution of the warming, particularly over recent decades when aerosol and land use—uncertain forcings with uncertain and potentially large efficacies—are thought to be relatively stable. This stability may allow the warming to be attributed to greenhouse gas and natural forcings which do not have as large uncertainty (Gregory & Forster, 2008). GFDL-CM4.0 has a GHG+NAT efficacy of close to 1, allowing direct implications about CO_2 sensitivity to be made. However, this may not be generally true of other climate models. Hansen et al. (2005) and Yoshimori and Broccoli (2008) cite efficacies for the natural solar and volcanic forcings near 0.9. Gillett and Arora (2013) show that historical GHG-forced warming and TCR are correlated in the CMIP5 models but that there is considerable variability in the relationship, especially when ozone forcing is included. Even in the case where other greenhouse gases have nonunitary efficacy, a finding of oversensitivity to these forcings provides a useful sensitivity constraint for climate projections because non-CO_2 gases will also contribute significantly to future warming, although less relative to CO_2 over time.

Figure 9 shows time series of the observed and CM4.0-simulated warming over the historical period. There is a significant discrepancy over the last 50 years of the experiment with CM4.0 producing a larger rate of warming than seen in the observations. CM4.0's responses to GHG+NAT and AERO+LU forcings are also shown in the figure along with their respective forcings scaled by the product of their efficacies and the transient sensitivity from the CO_2-only calculation. Because of the lack of recalcitrant warming over the historical experiment in CM4.0, forcing should be related to the contemporaneous response. Figure 9 shows that the responses to the forcing components conform reasonably well to those expected from the respective forcing time series supporting the interpretation of the partial forcing historical time series as radiative responses.

It is notable in Figure 9 that the response to AERO+LU forcing does not change over the 40-year period from about 1975 to 2014. Therefore, CM4.0's warming over this period is entirely due to GHG+NAT forcing. If the same is true of the observed climate, then the model must be too sensitive to these forcings because its post-1975 warming rate is considerably larger than observed. If, further, nature like CM4.0 has an efficacy of 1 for GHG+NAT forcing, then CM4.0 must also have excessive transient sensitivity to CO_2 in the historical period. Under these two assumptions, CM4.0's late-warming discrepancy is a manifestation of excessive transient sensitivity and would be expected to produce excessive climate change in future projections. Over the historical period as a whole, the oversensitivity to CO_2, other greenhouse gases, and natural forcings must be counteracted by an oversimulation of the cooling effects of land use and aerosols in order to achieve CM4.0's agreement with the overall observed warming, albeit with a low bias in the time series average warming.

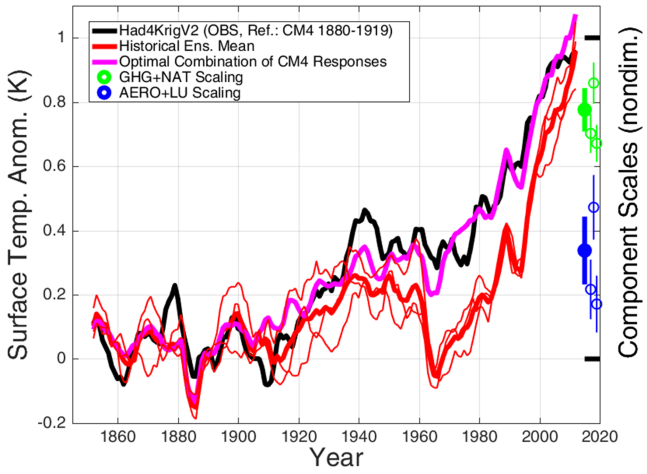

Figure 10. Observed and simulated global warming and scaling factors obtained by regressing observed temperature change onto CM4.0's greenhouse gas plus natural and aerosol plus land use responses with 95% confidence intervals (at right). All series are 5-year smoothed. The scaling factors using the historical ensemble mean (bold) and using each member individually to derive the aerosol plus land use component are shown. The resulting fit (purple) matches the observed temperature time series (black) quite well.

This line of reasoning can be formalized by using detection and attribution methods which project observed changes onto spatiotemporal patterns of response produced by climate models. The scaling factors produced this way detect a forced response when their ranges do not include 0 but also indicate an incompatibility between the simulated and observed response when their uncertainty ranges do not include 1. Gillett et al. (2012) performed detection and attribution analyses with varying levels of sophistication on CanESM2, a sensitive model with a TCR of 2.3 K. They found that, when the period from 1850 to 2010 was used, all methods gave similar scalings and showed that the model's responses to greenhouse gas and aerosol forcings were too large since the scaling factors were significantly less than 1. The simplest method used was an ordinary least squares regression of observed global mean temperature onto the model responses to individual forcings.

Here we apply this simple ordinary least squares regression method to CM4.0, projecting the Had4KrigV2 global warming onto CM4.0's 5-year smoothed GHG+NAT and AERO+LU responses obtaining scaling factors of 0.78 and 0.34, respectively. The fit to the observed global warming after applying these scaling factors to the model response components is very good ($r = 0.96$). The scaling factors are significantly less than 1 (95% confidence). We find therefore, similarly to Gillett et al. (2012), that the response to both forcing components is too large (Figure 10). This is the case whether we use the ensemble mean historical temperature to construct the AERO+LU response or any of the three individual ensemble members (the GHG+NAT response is determined from a single experiment). We only needed to perform a single all-forced and GHG+NAT-forced historical experiments, in addition to the piControl experiment, to infer excessive transient sensitivity with this simple method.

Applying CM4.0's GHG+NAT scaling factor to its TCR gives a value of 1.6 K as an estimate of the true value. However, Gillett et al. (2012) recommend against interpreting the scaling factors from a single model this way. Although our result shows that another high sensitivity model has oversensitivity to forcing components in the historical period, important uncertainties are not fully sampled. To draw inferences about Earth's sensitivity, it is recommended to take a multimodel approach as in the upcoming CMIP6 DAMIP experiment (Gillett et al., 2016). Nevertheless, this exercise has shown that CM4.0's sensitivity is, at minimum, inconsistent with observations given the truth of two other aspects of its simulation: recent stability of AERO+LU forcing and unitary efficacy of GHG+NAT forcing.

It would not be possible to achieve the good fit shown in Figure 10 by reducing the sensitivity alone, which would be equivalent to using the same, less than 1, scaling factor for both the GHG+NAT and AERO+LU responses. Doing this would reduce the overall preindustrial to present-day warming to well below that observed. By reducing the scaling factor on the cooling AERO+LU response more than for GHG+NAT, the overall warming is brought back up toward that observed while improving the shape of the warming evolution. The extra reduction in AERO+LU response can be achieved either by reducing the radiative forcing or the efficacy of the radiative forcing. The product of these two needs to be reduced by a little more than 50% to achieve the best fit.

To this point we have assumed that the model's late-warming discrepancy is due to an error in the forced response rather than from a contribution of unforced warming, most likely to the observed temperature record. The confidence intervals on the scaling factors in Figure 10 are evidence against a natural variability explanation but are generated by the regression and reflect only the uncertainty represented by the high frequency variation of the fit residuals. An estimate of the uncertainty generated by low frequency internal variability comes from the three-member ensemble and the figure shows that none of these have large enough variations to account for the observed temperature record. Figure 11 gives a broader perspective on the potential for natural fluctuations to account for CM4.0's long temperature discrepancy between the 1950s and near present by including the eight 60-year averages available from the 500-year preindustrial control experiment. These temperature variations are also seen to be insufficient to account for the discrepancy.

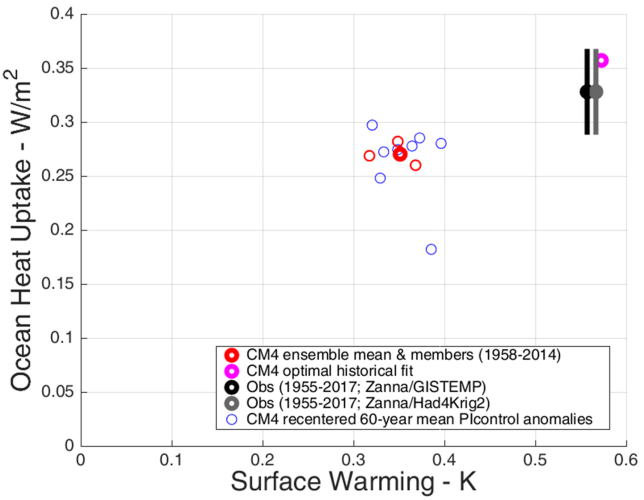

Figure 11. Gregory diagram for mid-1950s to near-present average warming and heat uptake. The experiment scalings shown in Figure 10 are applied to both surface temperature and heat uptake to obtain a hypothetical reduced sensitivity/aerosol+land use forcing response (pink) that agrees better with the observations (black and gray) than does the unadjusted CM4 (red). The PI control anomalies (blue) are the eight 60-year averages of global surface temperature and ocean heat uptake recentered on the historical ensemble mean values. Linear fit residuals are used in these averages in order to remove PI control drift. CM4's internal variations in temperature and heat uptake are negatively correlated ($\rho = -0.33$).

Several studies have attempted to explain recent discrepancies between CMIP5 model and observed warming by identifying warming pattern differences associated with either natural variability (e.g., Meehl et al., 2014) or missing forced responses (e.g., Seager et al., 2019) but the large scale ENSO-like pattern that has been identified is amenable to either explanation (Gregory et al., 2019). Gillett et al. (2012) showed that pattern information did not alter detection/attribution forcing factors or their uncertainties significantly relative to an analysis using only the evolution of global mean temperature when the longest observational period was used. Instead of using pattern information, to strengthen our identification of CM4.0's discrepancy as an error in its forced response we include ocean heat uptake observations in the comparison, hoping to take advantage of the strong association of surface warming and heat uptake in forced responses that is evident in Gregory diagrams (e.g., Figure 2).

We use the heat uptake observation from Zanna et al. (2019). This estimate agrees with available instrumental records but also extends back in time using Green's functions fit to a data-assimilation-constrained ocean circulation. The average temperature/heat uptake pairs are plotted on a type of Gregory diagram in Figure 11 showing that the low bias in model temperature is associated with a low bias in ocean heat uptake over the period since the mid 1950s in which there is a large discrepancy in the evolution of global mean temperature between CM4.0 and observations. The ocean heat uptake bias is relatively smaller because CM4.0's ocean heat uptake efficiency, the slope of the line from the origin, is larger than observed. We also apply the GHG+NAT and AERO+LU scalings derived earlier to surface temperature and heat uptake, showing a climate state for a hypothetical model obtained by reducing CM4.0's sensitivity and AERO+LU forcing. This state is in much better agreement with the observed state than any of the CM4 historical ensemble members or the hypothetical ensemble obtained by using the PI control natural fluctuations. The negative correlation of natural fluctuations in surface temperature and ocean heat uptake makes it more difficult for them to account for the model discrepancy where both biases have the same sign. The one (hypothetical) CM4.0 fluctuation that falls within the observed heat uptake confidence interval is associated with a surface cool anomaly and hence, has a larger surface temperature discrepancy with the observations.

Therefore, our interpretation of the ocean heat uptake comparison is that CM4.0's late-warming discrepancy is due to errors in its forced response, in particular that both its transient sensitivity and its product of AERO+LU forcing and forcing efficacy are too large. The warming prior to 1980 is too small because of excessive cooling from increasing AERO+LU forcing. After 1980, the warming is too large because the model is too sensitive and there is no longer a canceling influence from the AERO+LU forcing. In the remainder of this section we briefly examine two risks to the conclusion that CM4.0's TCR is too high: (1) The shape of the AERO+LU forcing may be incorrectly simulated by CM4.0, and (2) there may be a systematic bias in the volcanic forcing that produces an excessive recent warming trend in CM4.0.

First we look at the possibility that an incorrect simulation of the shape of the AERO+LU response prevents the model from simulating the historical warming correctly with its high TCR. We note that the shape of the response is similar to that of the forcing, which also does not have a significant trend over 1975–2014 (Figure 9). Figure 12a shows the response that would be needed for CM4.0 to agree with the observed warming without an adjustment to its GHG+NAT sensitivity obtained by subtracting the observed warming from CM4.0's GHG+NAT response. The required AERO+LU forced cooling increases steadily from 1960 to present day.

A uniform increase in aerosol forcing over this period is unlikely given that global emissions of SO_2, precursor to the main aerosol cooling agent, peaked around 1980 and have been falling since (Figure 12b; (Klimont et al., 2013; Smith et al., 2001). Myhre et al. (2017) show aerosol forcing for five models forced with CMIP6 aerosol precursor emissions. All five have either flat or increasing (less cooling) radiative forcing from 1990 to 2015.

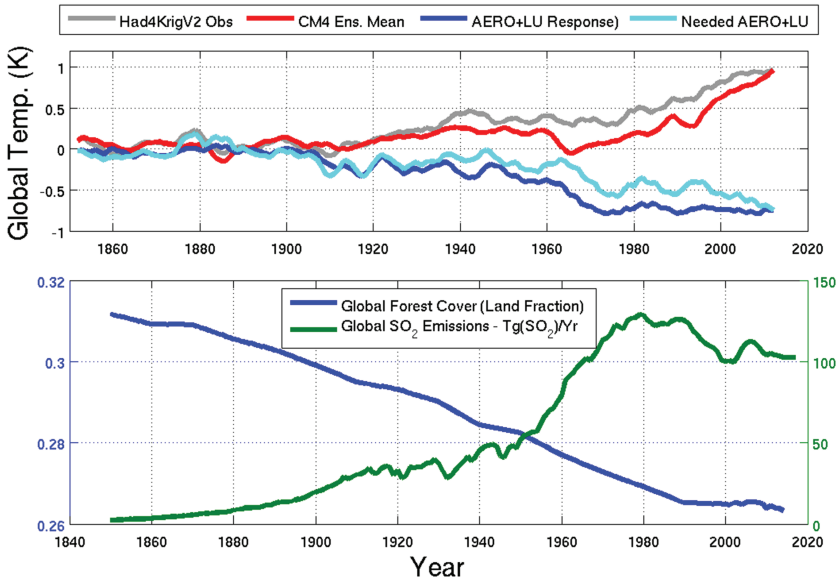

Figure 12. (a) CM4.0 aerosol plus land use response (blue) and the response needed to eliminate the discrepancy with the observed warming (light blue). All series are 5-year smoothed. (b) Two drivers of aerosol plus land use radiative forcing, global forest cover (blue), and SO_2 emissions (green) have stabilized in recent decades.

Land use forcing is also unlikely to produce a steady increase in AERO+LU cooling. Figure 12b shows global forest land area fraction from the CMIP6 forcing data set. The fall in area is abruptly reduced around 1990 while the required AERO+LU forcing in Figure 12a continues falling. For both aerosol and land use forcing, the basic drivers of forcing increase very little over the past several decades. Therefore, it is unlikely that the discrepancy between the observed and CM4 simulated warming over this period can be accounted for by a discrepancy in the shape of the AERO+LU forcing.

Finally, we examine the possibility that a systematic bias in the volcanic forcing might produce the late-warming discrepancy relative to observations. A series of strong volcanic eruptions from Agung in 1963 to Pinatubo in 1991 seem well placed to delay warming in response to greenhouse gas forcing which

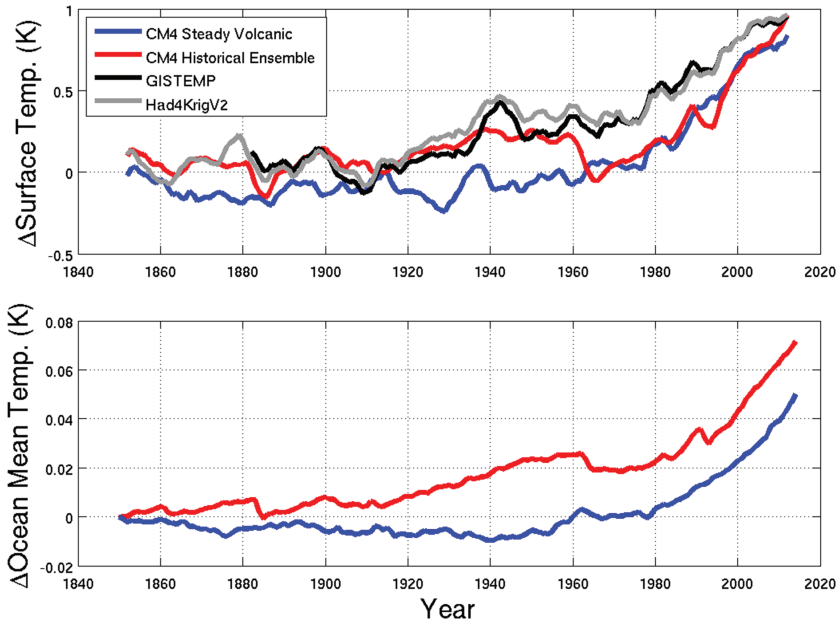

Figure 13. (a) Global surface temperature change (5-year smoothed) and (b) ocean average temperature change with time average volcanic forcing replacing time-varying forcing in a historical run (blue).

accelerated during the 1960's. In the CMIP6 protocol, the piControl experiment is forced with the historical period time mean volcanic aerosols. To explore the role of time-varying volcanic forcing, we run a historical experiment that retains the time average forcing used in the piControl experiment, in place of the time-varying forcing typically used in historical runs. This run is unexpectedly cooler in the time average than the standard historical and also has less heat uptake, the signature of a difference in radiative forcing (Figure 13). This appears to be due to a sensitive response to a saturation of aerosol effects during strong volcanic events in the time-varying forcing. There is an increase of historical period average aerosol direct negative forcing of 0.05 Wm^{-2} with the time mean volcanic forcing. This results in a cooling of 0.12 K, more than 4 times larger than expected from the model's TCR. Nevertheless, the forcing fluctuations are an order of magnitude larger than the time average forcing difference and yet the overall late-warming shape of surface and ocean average temperature is not significantly reduced by eliminating the time variation of volcanic forcing. The 1980–2014 surface warming trend with time mean volcanic forcing is 12% less than the CM4.0 all-forcing ensemble but still 50% and 30% greater than the trends in the GISTEMP and Had4KrigV2 observations, respectively. Therefore, it is unlikely that a systematic high bias in volcanic forcing could contribute significantly to CM4.0's late-warming discrepancy. However, CM4.0s simulated drop in temperature after 1960 that is not seen in the observations does seem attributable to the active volcanic forcing between 1963–1991. Also, CM4's gradual early 20th century warming consistent with observations seems in part attributable to a lack of strong volcanic forcing during that period. Therefore, these features might be affected by biases in volcanic forcing.

5. Summary and Discussion

CM4.0 was developed with the usual goals of improving model resolution, process fidelity, and simulation quality relative to the previous generation of GFDL models. But this model development cycle, unlike previous cycles, explicitly included consistent historical trends in the last category, additional to the quality of the present-day climatology and variability. There was an attempt to reduce the model's climate sensitivity in order to achieve this, primarily through cloud parameterization tuning (Zhao et al., 2018b) but also by increasing ocean deep-water ventilation which has been shown to reduce sensitivity (He et al., 2017; Winton et al., 2014). The effort to reduce the model's sensitivity during development has been justified ex post by our finding in this study that, even after reduction, CM4.0's TCR of 2.05 K remains likely too high. This result is obtained by using either direct reasoning from partial forcing experiments or by applying a simple detection attribution regression to those experiments and examining the scaling factors.

In particular, we showed that CM4.0's late-warming pattern of historical global warming indicates excessive transient sensitivity. This is the case even though CM4.0's overall preindustrial to present-day warming, heat uptake, and radiative forcing are consistent with observations and even though energy budget method sensitivity estimates based on CM4.0's present-day simulation agree with those based on present-day observations. We have shown that these energy budget constraints are ineffective for the model mainly because its large aerosol plus land use forcing efficacy is able to counteract the influence of high transient sensitivity on preindustrial to present-day changes. Because this efficacy is generally larger in CM4.0 than in other models, we expect that similar perfect model tests of the energy budget method in other models would identify errors that are smaller but also in the sense of underestimating the true sensitivities. We note that coupled experiments are needed to assess model efficacies and that observational constraints on these would likely need to come from observing the response of the coupled system making their constraint by observations difficult.

By examining CM4.0's simulation of historical ocean heat uptake and its response to aerosol, land use, and volcanic forcings, we showed that our inference of excessive sensitivity in CM4.0 has some robustness to assumptions made about the role of natural variability and evolution of forcings. Therefore, we might expect the positive relationship between TCR and recent warming trend to generalize to other models. A good test of this proposition will come from examining the upcoming CMIP6 model ensemble because it will cover the longest historical period with the best estimates of radiative forcings. To make a preliminary test of our inference of oversensitivity from excessive warming in recent decades, we can look for a relationship between the two in the older CMIP5 model ensemble, extending their historical simulations from 2006 to 2018 using the early years of the RCP4.5 projection. Likewise the CM4.0 historical has been extended with the early years of its SSP4.5 scenario run. Fyfe et al. (2013) and Rosenblum and Eisenman (2017) have shown

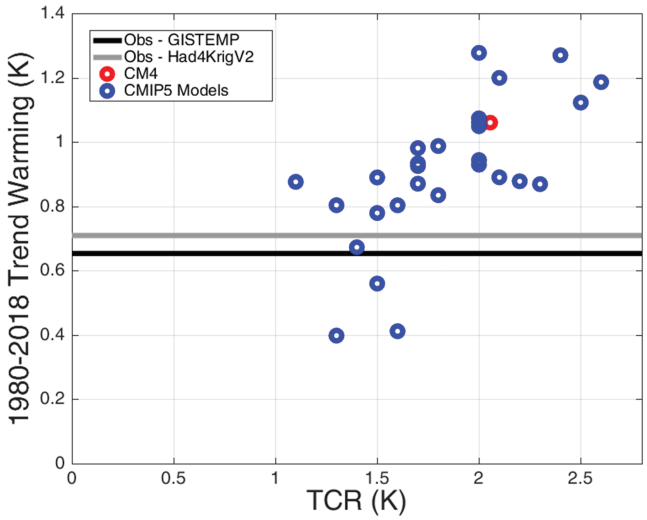

Figure 14. Historical warming trends plotted against TCR for CM4.0 (red) and CMIP5 model ensemble averages (blue, ensemble sizes in parentheses): ACCESS1-0 (1), ACCESS1-3 (1), bcc-csm1-1 (1), BCC-CSM1-1-m (1), BNU-ESM (1), CanESM2 (5), CCSM4 (6), CESM1-BGC (1), CESM1-CAM5 (3), CNRM-CM5 (1), CSIRO-Mk3-6-0 (9), GFDL-CM3 (1), GFDL-ESM2G (1), GFDL-ESM2M (1), GISS-E2-H (3), GISS-E2-R (6), HadGEM2-ES (3), INM-CM4 (1), IPSL-CM5A-LR (4), IPSL-CM5A-MR (1), IPSL-CM5B-LR (1), MIROC5 (5), MIROC-ESM (1), MPI-ESM-LR (3), MPI-ESM-MR (3), MRI-CGCM3 (1), NorESM1-ME (1), and NorESM1-M (1). Control experiment trends corresponding to 1850–2050 have been removed. Observed trends are also shown.

that CMIP models simulate excessive warming trends over recent decades. Figure 14 shows, in addition, that there is a relationship between the recent warming trend and TCR among the CMIP5 models (correlation = 0.67), and that only models with lower than average TCRs are able to reproduce the low trend seen in the observations. The aerosol forcing has been updated for CMIP6 (Myhre et al., 2017) and the new forcing has a flatter shape over recent decades which may sharpen the relationship.

The two sensitivity metrics that characterize the ocean's role in slowing radiatively forced warming, the ocean heat uptake efficiency and efficacy, are larger in CM4.0 than for the average climate model indicating that the model's above average sensitivity does not stem from its ocean simulation. However, AM4.0's Cess feedback—the top-of-atmosphere radiative flux in response to a uniform SST warming—is between that of AM2 and AM3, but closer to AM2, while the estimated ECS for CM4.0 is somewhat larger than that of CM3 and much larger than that of CM2-based models (Zhao et al., 2018b). However, the Cess feedback may underrepresent changes in high latitude feedbacks. So far CM4.0's high sensitivity has not been traced to the atmospheric component either. Since high transient sensitivity reduces the quality of CM4.0's historical simulation and will likely bias its projections, we regard its attribution as a high priority for future work.

Looking beyond the reasons for high model sensitivity to its implications, we have shown that GFDL-CM4.0 joins CanESM2 (Gillett et al., 2012) as a high sensitivity model that has been identified as too sensitive by a simple detection and attribution exercise using global mean temperature. We have also shown, more generally, that high sensitivity models overestimate the warming trend in recent decades when aerosol and land use forcing changes only weakly offset increasing greenhouse gas forcing. Therefore, it may be that transient sensitivities of 2 K or larger will be judged unlikely based on a broader evaluation of CMIP6 model historical simulations. This would be good news for efforts to keep global warming below target values because climate model ensembles are used to formulate the emissions limits for these targets and a reliance on lower sensitivity models for this purpose will lead to higher emissions limits.

Acknowledgments

We thank the GFDL community for the sustained effort that produced GFDL-CM4.0. Nadir Jeevanjee, Yi Ming, Fabien Paulot, and two anonymous reviewers are thanked for helpful comments on the manuscript. GFDL-CM4.0 data are available in the CMIP6 archive (Guo et al., 2018). Data used for this study are available online (at ftp://nomads.gfdl.noaa.gov/users/Mike.Winton). Alistair Adcroft and Levi Silvers were supported by Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

References

Andrews, T., Gregory, J., Paynter, D., Silvers, L., Zhou, C., Mauritsen, T., et al. (2018). Accounting for changing temperature patterns increases historical estimates of climate sensitivity. *Geophysical Research Letters*, *45*, 8490–8499. https://doi.org/10.1029/2018GL078887

Andrews, T., Gregory, J., & Webb, M. (2015). The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *Journal of Climate*, *28*, 1630–1648. https://doi.org/10.1175/JCLI-D-14-00545.1

Andrews, T., Gregory, J. M., Webb, M. J., & Taylor, K. E. (2012). Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophysical Research Letters*, *39*, L09712. https://doi.org/10.1029/2012GL051607

Armour, K. C. (2017). Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. *Nature Climate Change*, *7*, 331–335.

Armour, K. C., Bitz, C. M., & Roe, G. H. (2013). Time-varying climate sensitivity from regional feedbacks. *Journal of Climate*, *26*, 4518–4534. https://doi.org/10.1175/JCLI?D?12?00544.1

Chung, E., & Soden, B. (2015). An assessment of methods for computing radiative forcing in climate models. *Environmental Research Letters*, *10*, 074004.

Cowtan, K., & Way, R. (2014). Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1935–1944.

Davin, E. L., de Noblet-Ducoudr, N., & Friedlingstein, P. (2007). Impact of land cover change on surface climate: Relevance of the radiative forcing concept. *Geophysical Research Letters*, *34*, L13702. https://doi.org/10.1029/2007GL029678

Fyfe, J., Gillett, N., & Zwiers, F. (2013). Overestimated global warming over the past 20 years. *Nature Climate Change*, *3*, 767–769.

Gillett, N. P., & Arora, A. K. (2013). Constraining the ratio of global warming to cumulative CO_2 emissions using CMIP5 simulations. *Journal of Climate*, *26*, 6844–6858. https://doi.org/10.1175/JCLI-D-12-00476.s1

Gillett, N. P., Arora, V. K., Flato, G. M., Scinocca, J. F., & von Salzen, K. (2012). Improved constraints on 21st-century warming derived using 160 years of temperature observations. *Geophysical Research Letters*, *39*, L01704. https://doi.org/10.1029/2011GL050226

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., et al. (2016). The detection and attribution model intercomparison project (DAMIP v1.0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697. https://doi.org/10.5194/gmd-9-3685-2016

Gregory, J. M., & Andrews, T. (2016). Variation in climate sensitivity and feedback parameters during the historical period. *Geophysical Research Letters*, *43*, 3911–3920. https://doi.org/10.1002/2016GL068406

Gregory, J. M., Andrews, T., Ceppi, P., Mauritsen, T., & Webb, M. J. (2019). How accurately can the climate sensitivity to CO_2 be estimated from historical climate change? *Climate Dynamics*. https://doi.org/10.1007/s00382-019-04991-y

Gregory, J. M., Andrews, T., & Good, P. (2015). The inconstancy of the transient climate response parameter under increasing CO_2. *Philosophical Transactions of the Royal Society*, *373*. https://doi.org/10.1098/rsta.2014.0417

Gregory, J. M., & Forster, P. M. (2008). Transient climate response estimated from radiative forcing and observed temperature change. *Journal of Geophysical Research*, *113*, D23105. https://doi.org/10.1029/2008JD010405

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., et al. (2004). A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical Research Letters*, *31*, L03205. https://doi.org/10.1029/2003GL018747

Guo, H., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., et al. (2018). NOAA-GFDL GFDL-CM4 model output. https://doi.org/10.22033/ESGF/CMIP6.1402.

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., et al. (2005). Efficacy of climate forcings. *Journal of Geophysical Research*, *110*, D18104. https://doi.org/10.1029/2005JD005776

He, J., Winton, M., Vecchi, G., Jia, L., & Rugenstein, M. (2017). Transient climate sensitivity depends on base climate ocean circulation. *Journal of Climate*, *30*(4), 1493–1504. https://doi.org/10.1175/JCLI-D-16-0581.1

Held, I., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*, 3691–3727. https://doi.org/10.1029/2019MS001829

Held, I., Winton, M., Takahashi, K., Delworth, T., Zeng, F., & Vallis, G. (2010). Probing the fast and slow components of global warming by returning abruptly to pre-industrial forcing. *Journal of Climate*, *23*(9). https://doi.org/10.1175/2009JCLI3466.1

Jones, A., Collins, W., & Torn, M. (2013). On the additivity of radiative forcing between land use change and greenhouse gases. *Geophysical Research Letters*, *40*, 4036–4041. https://doi.org/10.1002/grl.50754

Kiehl, J. (2007). Twentieth century climate model response and climate sensitivity. *Geophysical Research Letters*, *34*, L22710. https://doi.org/10.1029/2007GL031383

Klimont, Z., Smith, S. J., & Cofala, J. (2013). The last decade of global anthropogenic sulfur dioxide: 2000–2011 emissions. *Environmental Research Letters*, *8*(1), 014003.

Kuhlbrodt, T., & Gregory, J. (2012). Ocean heat uptake and its consequences for the magnitude of sea level rise and climate change. *Geophysical Research Letters*, *39*, L18608. https://doi.org/10.1029/2012GL052952

Lewis, N., & Curry, J. (2015). The implications for climate sensitivity of AR5 forcing and heat uptake estimates. *Climate Dynamics*, *45*(3-4), 1009–1023.

Lewis, N., & Curry, J. (2018). The impact of recent forcing and ocean heat uptake data on estimates of climate sensitivity. *Journal of Climate*, *31*, 6051–6071. https://doi.org/10.1175/JCLI-D17-0667.1

Li, C., von Storch, J., & Marotzke, J. (2013). Deep-ocean heat uptake and equilibrium climate response. *Climate Dynamics*, *40*, 1071–1086. https://doi.org/10.1007/s00382-012-1350-z

Marvel, K., Schmidt, G., Miller, R., & Nazarenko, L. (2015). Implications for climate sensitivity from the response to individual forcings. *Nature Climate Change*, *6*, 386–389.

Meehl, G., Teng, H., & Arblaster, J. (2014). Climate model simulations of the observed early-2000s hiatus of global warming. *Nature Climate Change*, *4*, 898–902. https://doi.org/10.1038/nclimate2357

Murphy, J. M. (1995). Transient response of the Hadley Centre coupled ocean-atmosphere model to increasing carbon dioxide: Part III. Analysis of global-mean response using simple models. *Journal of Climate*, *8*, 496–514.

Myhre, G., Aas, W., Cherian, R., Collins, W., Faluvegi, G., Flanner, M., et al. (2017). Multi-model simulations of aerosol and ozone radiative forcing due to anthropogenic emission changes during the period 1990–2015. *Atmospheric Chemistry and Physics*, *17*, 2709–2720.

Otto, A., Otto, F. E. L., Boucher, O., Church, J., Hegerl, G., Forster, P. M., et al. (2013). Energy budget constraints on climate response. *Nature Geoscience*, *1043*(6), 415–416.

Paynter, D., Frlicher, T., Horowitz, L., & Silvers, L. (2018). Equilibrium climate sensitivity obtained from multimillennial runs of two GFDL climate models. *Journal of Geophysical Research: Atmospheres*, *123*, 1921–1941. https://doi.org/10.1002/2017JD027885

Proistosescu, C., & Huybers, P. J. (2017). Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Science Advances*, *3*(7), e1602821. https://doi.org/10.1126/sciadv.1602821

Richardson, M., Cowtan, K., Hawkins, E., & Stolpe, M. (2016). Reconciled climate response estimates from climate models and the energy budget of Earth. *Nature Climate Change*, *6*, 931–935. https://doi.org/10.1038/nclimate3066

Rosenblum, E., & Eisenman, I. (2017). Sea ice trends in climate models only accurate in runs with biased global warming. *Journal of Climate*, *30*, 6265–6278. https://doi.org/10.1175/JCLI-D-16-0455.1

Rotstayn, L., Collier, M., Shindell, D., & Boucher, O. (2015). Why does aerosol forcing control historical global-mean surface temperature change in CMIP5 models? *Journal of Climate*, *28*, 6608–6625. https://doi.org/10.1175/JCLI-D-14-00712.1

Seager, R., Cane, M., Henderson, N., Lee, D.-E., Abernathy, R., & Zhang, H. (2019). Strengthening tropical pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nature Climate Change*, *9*, 517–522.

Senior, C. A., & Mitchell, J. F. B. (2000). The time-dependence of climate sensitivity. *Geophysical Research Letters*, *27*(17), 2685–2688. https://doi.org/10.1029/2000GL011373

Shindell, D. (2014). Inhomogeneous forcing and transient climate sensitivity. *Nature Climate Change*, *4*, 274–277. https://doi.org/10.1038/nclimate2136

Smith, S. J., Pitcher, H., & Wigley, T. (2001). Global and regional anthropogenic sulfur dioxide emissions. *Global and Planetary Change*, *29*(1), 99–119. https://doi.org/10.1016/S0921-8181(00)00057-6

Stevens, B., Sherwood, S., Bony, S., & Webb, M. (2016). Prospects for narrowing bounds on Earth's equilibrium climate sensitivity. *Earth's Future*, *4*, 512–522. https://doi.org/10.1002/2016EF000376

Tebaldi, C., & Arblaster, J. (2014). Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, *122*(3), 459–471. https://doi.org/10.1007/s10584-013-1032-9

Winton, M., Adcroft, A., Griffies, S., Hallberg, R., Horowitz, L., & Stouffer, R. (2013). Influence of ocean and atmosphere components on simulated climate sensitivities. *Journal of Climate*, *26*(1). https://doi.org/10.1175/JCLI-D-12-00121.1

Winton, M., Anderson, W., Delworth, T., Griffies, S., Hurlin, W., & Rosati, A. (2014). Has coarse ocean resolution biased simulations of transient climate sensitivity? *Geophysical Research Letters*, *41*, 8522–8529. https://doi.org/10.1002/2014GL061523

Winton, M., Takahashi, K., & Held, I. (2010). Importance of ocean heat uptake efficacy to transient climate change. *Journal of Climate*, *23*(9), 2333–2344. https://doi.org/10.1175/2009JCLI3139.1

Yoshimori, M., & Broccoli, A. (2008). Equilibrium response of an atmosphere-mixed layer ocean model to different radiative forcing agents: Global and zonal mean response. *Journal of Climate*, *21*, 4399–4423. https://doi.org/10.1175/2008JCLI2172.1

Zanna, S., Khatiwala, L., Gregory, J., Ison, J., & Heimbach, P. (2019). Global reconstruction of historical ocean heat storage and transport. *PNAS*, *116*(4), 1126–1131. https://doi.org/10.1073/pnas.1808838115

Zhao, M., Golaz, J.-C., Held, I., Guo, H., Balaji, V., Benson, R., et al. (2018a). The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, *10*, 691–734. https://doi.org/10.1002/2017MS001208

Zhao, M., Golaz, J.-C., Held, I., Guo, H., Balaji, V., Benson, R., et al. (2018b). The GFDL global atmosphere and land model AM4.0/LM4.0: 2. Model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, *10*, 735–769.

Zhou, C., Zelinka, M., & Klein, S. (2017). Analyzing the dependence of global cloud feedback on the spatial pattern of sea surface temperature change with a Green's function approach. *Journal of Advances in Modeling Earth Systems*, *9*, 2174–2189. https://doi.org/10.1002/2017MS001096