AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2016JD025320

Key Points:
- We recommend a protocol for estimating ERF in GCMs
- Error characteristics of ERF make diagnosing small forcings hard
- Some CMIP6 protocols may not work (AerCHemMIP in particular)

Supporting Information:
- Supporting Information S1

Correspondence to:
P.M. Forster,
p.m.forster@leeds.ac.uk

Citation:
Forster, P. M., T. Richardson, A. C. Maycock, C. J. Smith, B. H. Samset, G. Myhre, T. Andrews, R. Pincus, and M. Schulz (2016), Recommendations for diagnosing effective radiative forcing from climate models for CMIP6, *J. Geophys. Res. Atmos.*, *121*, 12,460–12,475, doi:10.1002/2016JD025320.

Received 4 MAY 2016
Accepted 6 OCT 2016
Accepted article online 8 OCT 2016
Published online 31 OCT 2016

Recommendations for diagnosing effective radiative forcing from climate models for CMIP6

Piers M. Forster[1], Thomas Richardson[1], Amanda C. Maycock[1], Christopher J. Smith[1], Bjorn H. Samset[2], Gunnar Myhre[2], Timothy Andrews[3], Robert Pincus[4], and Michael Schulz[5]

[1]University of Leeds, Leeds, UK, [2]CICERO, Oslo, Norway, [3]Met Office, Exeter, UK, [4]University of Colorado Boulder, Boulder, Colorado, U.S.A., [5]Norwegian Meteorological Institute, Oslo, Norway

Abstract The usefulness of previous Coupled Model Intercomparison Project (CMIP) exercises has been hampered by a lack of radiative forcing information. This has made it difficult to understand reasons for differences between model responses. Effective radiative forcing (ERF) is easier to diagnose than traditional radiative forcing in global climate models (GCMs) and is more representative of the eventual temperature response. Here we examine the different methods of computing ERF in two GCMs. We find that ERF computed from a fixed sea surface temperature (SST) method (ERF_fSST) has much more certainty than regression based methods. Thirty year integrations are sufficient to reduce the 5–95% confidence interval in global ERF_fSST to 0.1 W m^{-2}. For 2xCO2 ERF, 30 year integrations are needed to ensure that the signal is larger than the local confidence interval over more than 90% of the globe. Within the ERF_fSST method there are various options for prescribing SSTs and sea ice. We explore these and find that ERF is only weakly dependent on the methodological choices. Prescribing the monthly averaged seasonally varying model's preindustrial climatology is recommended for its smaller random error and easier implementation. As part of CMIP6, the Radiative Forcing Model Intercomparison Project (RFMIP) asks models to conduct 30 year ERF_fSST experiments using the model's own preindustrial climatology of SST and sea ice. The Aerosol and Chemistry Model Intercomparison Project (AerChemMIP) will also mainly use this approach. We propose this as a standard method for diagnosing ERF and recommend that it be used across the climate modeling community to aid future comparisons.

1. Introduction

The ubiquitous framework for understanding surface temperature changes due to specific radiative drivers is based around energy budget analyses that split forcing and response [*Boucher et al*., 2013; *Myhre et al*., 2013; *Sherwood et al*., 2015]. This framework has proved invaluable for characterizing the drivers of climate change and understanding many aspects of how the climate is expected to respond to human and natural drivers of change.

Forcing has traditionally been calculated as either an instantaneous radiative forcing or stratospherically adjusted radiative forcing, measured as a W m^{-2} change in irradiance at the tropopause for a trace gas change, land use change, solar irradiance change, or aerosol perturbation [*Ramaswamy et al*., 2001; *Sherwood et al*., 2015]. However, these traditional definitions of radiative forcings do not account for forcing-driven changes in cloudiness and other rapid adjustments that alter the global energy balance and ultimately affect the eventual climate response. Such rapid adjustments are accounted for within the concept of effective radiative forcing (ERF). ERF is defined as the irradiance change at the top of atmosphere (TOA) following a perturbation to the climate system taking into account any rapid adjustments. The exact definition of rapid adjustment and, therefore, ERF varies with the calculation method. ERF is more uncertain than traditional radiative forcing because it involves interactions with multiple aspects of the climate system, e.g., clouds, but has become the metric of choice [*Boucher et al*., 2013] because a forcing-feedback framework based on ERF gives a more complete picture of the overall expected energy budget change [*Chung and Soden*, 2015; *Sherwood et al*., 2015]. Climate sensitivity parameters (the degree of warming per unit forcing) are less dependent on the forcing agent when rapid adjustments are accounted for [*Hansen et al*., 2005; *Shine et al*., 2003]. ERF is also more readily calculable from standard climate model diagnostics [*Sherwood et al*., 2015].

A major disadvantage of ERF as a metric is that it depends on its method of calculation and there is, as yet, no agreed method. Two main methods are employed in models: ERF_fSST, which is diagnosed from fixed sea surface temperature and sea ice integrations [*Hansen et al*., 2005], or ERF_reg, which is diagnosed from regression of TOA irradiance against global surface temperature change in integrations of coupled ocean-atmosphere models where a forcing such as CO_2 is abruptly increased [*Gregory et al*., 2004]. An advantage of the ERF_fSST method over the ERF_reg method is that it can be employed to make a transient estimate of ERF from a scenario with time-varying forcings [*Andrews and Ringer*, 2014]. The ERF_fSST approach is similar to separating forcing and feedback by timescales, as ERF_fSST includes responses on the timescales of atmosphere and land surface change but not the responses associated with longer ocean timescales. A downside is that the land surface response is included in the estimate of global ERF_fSST, and this leads to a change of global mean surface temperature. A global mean temperature change makes it more difficult to fit ERF_fSST into a simple forcing-feedback framework whereby feedbacks are related to global temperature change and forcings are not [*Sherwood et al*., 2015]. Another downside is that rapid adjustments are no longer treated as a coupled problem, and at least in one model, the ocean plays an important role in governing their development [*Rugenstein et al*., 2016].

One potential disadvantage shared by ERF_fSST and ERF_reg is that both methods require custom model integrations. Alternatives exist but have greater weaknesses. ERF can be inferred from TOA irradiances in existing coupled climate model runs, for example, by assuming constant known climate feedbacks [*Forster et al*., 2013; *Forster and Taylor*, 2006] or using an impulse response model that assumes known time-varying feedbacks [*Larson and Portmann*, 2016]. Because these methods assume known feedbacks, it is difficult to attribute the intermodel spread in temperature change to differences in intermodel forcing and/or differences in feedback mechanisms. ERF may also be estimated by adding estimates of rapid adjustments onto instantaneous radiative forcing estimates [*Chung and Soden*, 2015], but this relies on the use of radiative kernels which introduces further uncertainty [*Chung and Soden*, 2015].

The lack of an agreed methodology is one reason why few of the world's climate models routinely report forcings. This has been a community issue for over a decade that has been particularly noticeable when trying to understand a model's response. Modeling groups do not typically know if their trace gas or aerosol change perturbation experiments give the radiative forcing expected from standard offline estimates of forcing [*Chung and Soden*, 2015; *Myhre et al*., 2013], nor do they know how their forcings compare to other models under standard historic and future scenarios [*Forster et al*., 2013].

Leaving forcings undiagnosed greatly limits the understanding of why the models get the responses they do and why they might differ [*Chung and Soden*, 2015; *Forster et al*., 2013; *Marvel et al*., 2015; *Shindell*, 2014; *Vial et al*., 2013]. One benefit of diagnosing historic forcing from models is that it will help us to resolve a key discrepancy over estimates of Equilibrium Climate Sensitivity (ECS) identified in the last Intergovernmental Panel on Climate Change (IPCC) report [*Intergovernmental Panel on Climate Change*, 2013]. ECS is defined as the global mean surface warming at equilibrium from a sustained doubling of carbon dioxide from preindustrial levels. In the IPCC report estimates of ECS based on historical energy budget analyses were centered around 2 K, consistently smaller than that derived from other methods, which were centered around 3 K [*Collins et al*., 2013]. The historic energy budget approach is beginning to be tested in models and such testing has led to important insights into why estimates may differ. Forcing driver efficacy [*Marvel et al*., 2015; *Shindell*, 2014], shifting spatial patterns of response [*Armour et al*., 2012; *Rose et al*., 2014] and temporal evolution of feedback [*Andrews et al*., 2015; *Gregory and Andrews*, 2016] have all been mooted as a possible cause of differences between ECS estimates. However, a lack of ERF estimates from multiple models means that perfect-model tests of how sensitivity has varied through time have not been possible and our knowledge is not as developed as it could be. Knowing historic forcing in models allows us to accurately diagnose how their feedbacks vary in space and time and allows one to assess how feedbacks for historical scenarios compare to feedbacks from idealized doubling carbon dioxide experiments. Not knowing historic ERF also makes it difficult to infer estimates of transient climate response [*Gregory and Forster*, 2008; *Storelvmo et al*., 2016], and challenging to understand causes of decadal variations in surface temperature [*Fyfe et al*., 2016; *Marotzke and Forster*, 2015], hampering our ability to attribute past climate trends to particular causes [*Stevens*, 2015].

In this paper we make recommendations for how ERFs can be routinely and consistently diagnosed across models to support Coupled Model Intercomparison Project Phase Six (CMIP6). We specifically test ERF

diagnosis methods employed as part of the Radiative Forcing and Aerosol-Chemistry MIPs associated with CMIP6 (RFMIP [*Pincus et al*., 2016 and AerChemMIP [*Collins et al*., 2016], respectively). We compare the uncertainty characteristics of ERF_fSST integrations with those of ERF_reg. We also compare methods for estimating the transient evolution of ERF (ERF_trans) that will be employed for both MIPs, discussing the pros and cons of the various methods.

2. Methods

We base most of the following analysis on integrations where step change perturbations have been made to different forcing agents such as CO_2 as part of the Precipitation Driver Response Model Intercomparison (PDRMIP). In this work we analyze data from the Hadley Centre Global Environmental Model (HadGEM2) and Community Earth System Model (CESM1) model integrations.

Five climate perturbation experiments were simulated: a doubling of CO_2 concentration (hereafter denoted *2xCO2*), tripling of CH_4 concentration (*3xCH4*), 2% increase in solar constant (*2%Sol*), 10 times BC concentration or emissions (*10xBC*), and 5 times SO_4 concentrations or emissions (*5xSul*).

These experiments were set up slightly differently in the two models which means that the computed forcings or responses are not expected to be quantitatively similar. For example, CESM1 scaled present-day aerosol concentrations based on AeroCom Phase II [see, e.g., *Samset et al*., 2013], whereas HadGEM2 scaled preindustrial emissions. For the regression analyses HadGEM2 employed a full ocean model, whereas CESM1 employed a simple slab-ocean model.

For the baseline and each perturbation experiment, the models ran two sets of simulations: one keeping sea surface temperatures and sea ice fixed (hereafter denoted *fSST*) and one with a coupled ocean (*coupled*). The fSST simulations were run for 30 years and used to compute ERF_fSST, and the coupled simulations for 100 years to compute ERF_reg. For the regional analyses, all model data were regridded from their native resolutions (1.875° × 1.25° grid boxes in HadGEM2 and 2.5° × 2° grid boxes in CESM1) to 1° × 1° resolution before comparisons were made. The fSST-type simulations adopted the sstClim methodology from CMIP5, where the SST and sea ice climatology was based on an annually repeating monthly averaged preindustrial climatology of SST and sea ice fraction was interpolated to give daily boundary conditions.

2.1. ERF_reg

Global mean ERF_reg was computed by linearly regressing the global, annual mean net TOA flux change relative to the baseline simulation against the change in global mean surface air temperature (T) in the coupled simulations. ERF_reg is defined as the intercept of the regression line with the $T = 0$ line. Local ERF_reg was calculated by linearly regressing the local annual mean net TOA flux change against the change in global mean surface air temperature. The 5%–95% confidence intervals for ERF_reg were calculated using the t distribution as shown in equation (1),

$$\mathrm{CI}_{\mathrm{reg}} = \mathrm{SE}_{\mathrm{reg}} \times \ t_{\mathrm{value}} \tag{1}$$

where $\mathrm{CI}_{\mathrm{reg}}$ is the ERF_reg 5%–95% confidence interval, $\mathrm{SE}_{\mathrm{reg}}$ is the standard error of the regression intercept, and t_{value} is the t value from the t distribution with degrees of freedom given by the number of years regressed over minus two. The standard error of the regression intercept was calculated using

$$\mathrm{SE}_{\mathrm{reg}} = \sqrt{\frac{\Sigma T_i{}^2 \Sigma (y_i - \overline{y}_I)^2}{n(n-2)\Sigma \left(T_i - \overline{T}_I\right)^2}}, \tag{2}$$

where T_i is the annual average temperature change, y_i is the annual average net TOA flux, n is the number of years regressing over, and overbars denote the average value of that quantity. We also employed multiple ensemble members to compute ERF_reg. To test uncertainty in the regression approaches, we ran five more ensemble members out to 20 years for 2xCO2 in the CESM1 model. The annually averaged data from the extra ensemble members were simply added to the number of points used in the regression analysis calculation of the intercept and 5%–95% confidence intervals.

2.2. ERF_fSST

ERF_fSST was taken as the difference in global mean net TOA flux between the perturbed and control fSST simulations [e.g., *Hansen et al*., 2005]. Local differences were integrated and area weighted to derive

a global mean ERF_fSST. The 5%–95% confidence intervals for ERF_fSST were calculated using the *t* distribution

$$CI_{fSST} = SE_{fSST} \times t_{value} \quad (3)$$

where CI_{fSST} is the ERF_fSST 5%–95% confidence interval, SE_{fSST} is the ERF_fSST standard error, and t_{value} is the *t* value from the *t* distribution with degrees of freedom given by the length of integration (in years) minus one. The standard error in ERF_fSST was calculated using equation (4):

$$SE_{fSST} = \frac{\sigma}{\sqrt{n}}, \quad (4)$$

where σ is the standard deviation of the annual mean anomaly in TOA radiation and *n* is the number of years in the integration.

2.3. IRF

Instantaneous radiative forcings (IRFs) were also diagnosed using a double-call methodology in which every call to the radiation scheme was repeated returning the trace gas or aerosol concentration to its preindustrial value [*Chung and Soden*, 2015]. CESM1 computed IRF as the difference in net downwelling flux at TOA from the first five years from the fSST integration. HadGEM2 computed IRF as the difference in net flux at the model level corresponding to the WMO definition of the tropopause, where lapse rate falls to 2 K km^{-1} or less, from years two and three of a 3 year fSST integration. These differences were computed locally and these were integrated and area weighted to derive a global mean IRF. CESM1 only computed IRF for aerosol changes. We chose to report IRF at the tropopause level in HadGEM2 as we also made estimates of greenhouse gas IRF in this model. By definition, stratospheric adjustments are not included in the IRF estimate. For greenhouse gas perturbations it is important to include stratospheric adjustment in the total radiative forcing estimate and recording fluxes at the tropopause makes stratospheric adjustment a part of the total rapid adjustment in a consistent way, as the difference between IRF and ERF.

2.4. ERF_nudge

Another way to reduce interannual variability in the TOA imbalance is to combine the ERF_fSST approach with nudging techniques [*Kooperman et al*., 2012] that constrain the model evolution by relaxing the model toward a specified time-dependent dynamical state [*Telford et al*., 2008]. The prescribed conditions are taken from the ERA-Interim reanalysis [*Dee et al*., 2011], however an alternative is to use a preindustrial control run of the same model [*Kooperman et al*., 2012]. For ERF estimates atmospheric wind fields are typically nudged but temperatures are not nudged to allow the atmospheric temperature and cloud fields to rapidly adjust to the forcing. We perform such experiments in HadGEM2, where atmospheric winds are nudged to ERA-Interim values and atmospheric temperatures allowed to evolve from 2001 onward (ERF_nudge). Following the recommendation in *Telford et al*. [2008], nudging is not applied in the boundary layer to avoid instabilities arising from differences between the HadGEM2 and ERA-Interim model orography, or near the top of the model. The relaxation parameter is a subjective choice that drives the strength of the nudging, and is set to the inverse of the time step in the ERA-Interim data (21,600 s). In ERF_nudge experiments, the results from years two and three of a 3 year integration are used to determine ERF. Owing to the short simulation length it is not appropriate to evaluate uncertainty from these runs.

2.5. ERF_trans

In section 4 we also explore modifications of the ERF_fSST approach to provide transient estimates of ERF. We term these methods ERF_trans. We perform transient integrations in HadGEM2 and CESM1 to test ERF_trans approaches that have different methods of prescribing SSTs and sea ice. ERF_trans is defined as the difference in downwelling TOA flux between a run where all forcers have transient evolution and one where one or more components have emissions (or concentrations) fixed at preindustrial values. Both control and perturbation experiments use the same SST and sea ice fields. For some experiments SSTs and sea ice fields are taken from the same annually repeating climatology as used for ERF_fSST. This protocol, based on each model's own annually repeating preindustrial climatology for all its experiments, is adopted by RFMIP. Other experiments use evolving SST and sea ice fields, following the proposed AerChemMIP protocol [*Collins et al*., 2016] which prefers to use monthly averaged evolving SST and sea ice fields taken from a parallel Atmosphere-Ocean General Circulation Model (AOGCM) integration. In both cases, ERF_trans is defined

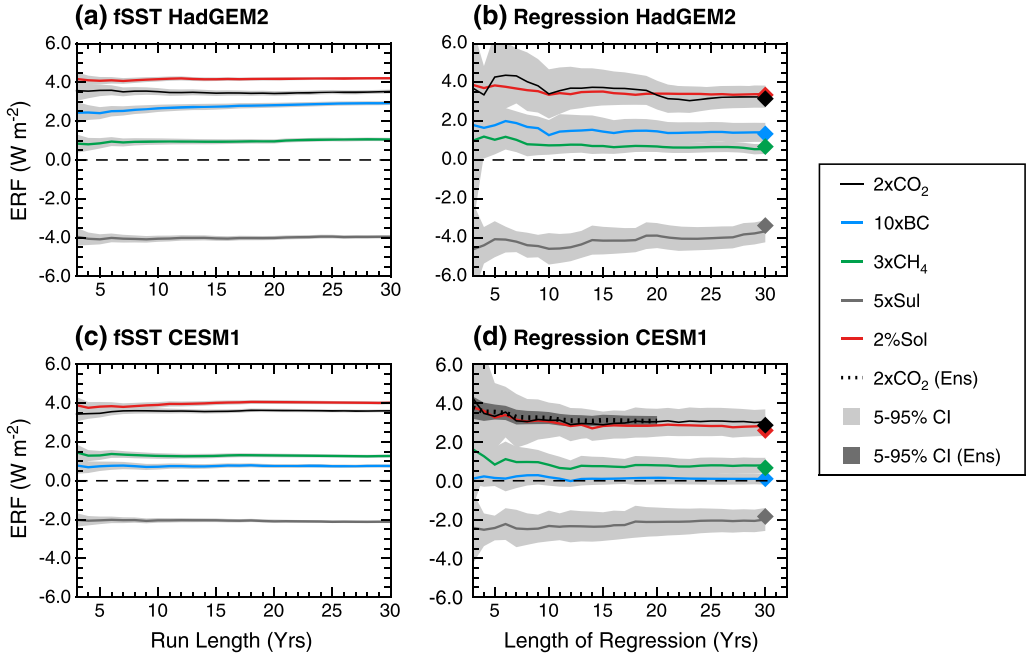

Figure 1. Global mean ERF (W m^{-2}) against (a and c) integration length for ERF_fSST and (b and d) regression length for ERF_reg. Results are presented for the five perturbation experiments implemented in HadGEM2 and CESM1. Grey shading denotes the 5%–95% confidence interval. Diamonds mark regression values after 100 years. Also shown as the dotted line on the CESM1 regression plot is the ERF diagnosed from regression of five ensemble members.

as the difference in downwelling TOA flux between a run where all forcers have transient evolution and one where one or more components have emissions (or concentrations) fixed at preindustrial values. ERF_trans could conceivably be defined as the difference in downwelling TOA flux between integrations where one forcing evolved compared to one where all forcings are fixed at preindustrial concentrations. However, such a method would not capture possible interdependence between forcing mechanisms (e.g., the dependence of CH_4 forcing on N_2O concentration) and would therefore give a less realistic estimate of ERF.

3. Comparing ERF_fSST and ERF_reg Approaches

This section explores how integration length and/or regression length affect the uncertainty characteristics of ERF diagnosed by the ERF_fSST and ERF_reg methods. Figure 1 compares the 5–95% confidence interval of the global ERF produced by the two methods and compares how ERF varies with integration length. For all forcings and in both models ERF_reg changes considerably as more years are included in the regression. This is due to nonlinearity in the relationship between TOA irradiance and surface temperature change and has been seen before for individual models and forcings [*Andrews et al.*, 2015; *Meraner et al.*, 2013]. This nonlinearity is likely caused by climate feedback variations over both time and space and leads to curvature in the relationship between TOA irradiance and surface temperature change. Uncertainty in ERF also reduce with the regression length, but they are always greater than 10% of the absolute ERF (for a single ensemble member). In contrast ERF_fSST is less dependent on run length and has consistently smaller uncertainty. The fSST control simulation which did not include ocean circulation-driven changes had similar variability to the coupled model control simulation indicating that the differences between ERF_fSST and ERF_reg confidence intervals were unlikely to be caused by the different ocean coupling in the two setups.

The choice of the number of years to integrate over for ERF_fSST, or to include in a regression analysis for ERF_reg, is somewhat subjective. For purposes of illustration we select 20 years for both. Figure 2 summarizes the 20 year results of Figure 1, presenting the uncertainty as error bars. The ERF_fSST and ERF_reg methods give more-or-less similar global ERFs for the different forcings, but there are also some interesting differences. In CESM1, ERF_fSST is more strongly positive than ERF_reg for all but 5xSul creating a systematic difference between the two approaches. The ERF patterns of the two approaches are shown and briefly discussed in the

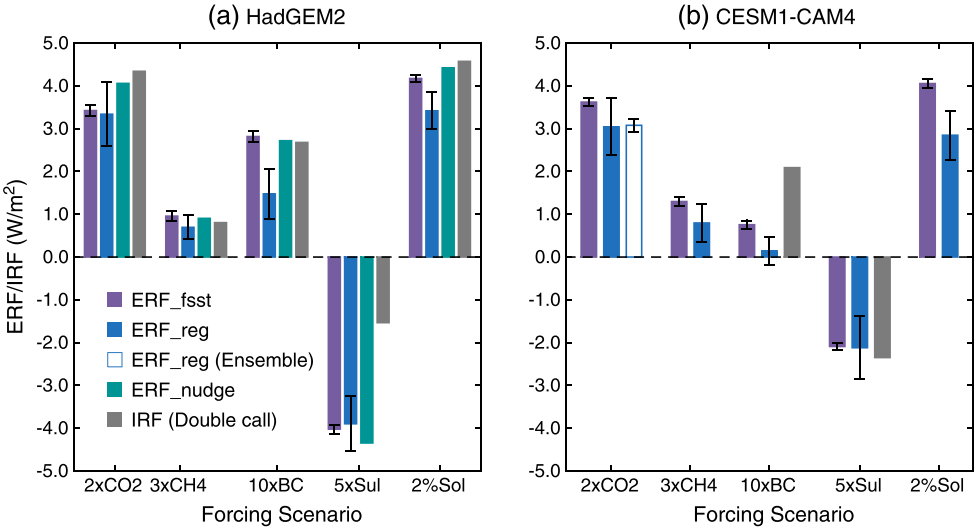

Figure 2. ERF values obtained by ERF_fSST and ERF_reg in the two models across the forcing experiments. Double-call IRF results for all experiments in HadGEM2 and for *10xBC* and *5xSul* in CESM1 are also given. ERF_nudge results are also shown for HadGEM2. Averaging lengths in ERF_fSST and regression lengths in ERF_reg are for 20 years. The results of ERF_reg for five ensemble members are also shown for *2xCO2* in CESM1. Error bars indicate the 5%–95% confidence interval. The confidence intervals are not evaluated for ERF_nudge and IRF.

supporting information. Patterns of ERF_fSST and ERF_reg are broadly similar to each other, but the ERF_reg pattern has larger uncertainty. Differences between the two ERF methods are expected. First, rapid adjustment estimates within the ERF_reg approach could have elements of SST pattern changes not captured by the ERF_fSST method [*Andrews et al.*, 2015]. Second, the land surface temperature change in the ERF_fSST approach contributes to a global mean temperature change and drives a TOA response. How rapid adjustments manifest themselves on this pattern of the land surface temperature response therefore affects the diagnosed forcing. A simple method of adjusting the ERF_fSST approach to account for its global surface temperature feedback has been tried before using an adjustment based on the global climate sensitivity parameter, but this method does not work particularly well in making estimates comparable [*Hansen et al.*, 2005; *Sherwood et al.*, 2015]. The differences in these processes would be expected to vary both by model and pattern of temperature response, making it difficult to derive a simple method of directly comparing the two types of ERF estimate. Indeed, a systematic difference between ERF_fSST and ERF_reg is not seen in HadGEM2.

Nevertheless, ERF_fSST has consistently smaller sampling errors than a single ensemble member for ERF_reg. Note that five ensemble members of the ERF_reg method are needed to give comparable confidence interval to ERF_fSST for the 2xCO2 case examined in CESM1 (Figure 2).

The difference in forcing between the ERF and the double-call IRF methods can be attributed to the effects of rapid adjustments to the land surface, troposphere, and stratosphere, which are not accounted for in the IRF. Examining Figure 2, there is evidence of a rapid adjustment with most forcing mechanisms in both models. The sign and proportional strength of this adjustment varies with forcing mechanism and model. A consistent feature in both models is that the ERF_reg from 10xBC is considerably smaller than its IRF, implying a negative forcing from rapid adjustment. There is evidence of strongly negative rapid adjustment for 5xSul in HadGEM2 but not in CESM1. Tropospheric rapid cloud adjustments are the likely cause of differences for 10xBC and 5xSul, in particular the semidirect effect for 10xBC and aerosol-cloud interaction for 5xSul [*Zelinka et al.*, 2014]. For 2xCO2, the all-sky stratospherically adjusted radiative forcing has been shown to be 14% lower than the IRF at the tropopause [*Myhre and Stordal*, 1997]. Stratospheric adjustment therefore likely explains part of the reason why the 2xCO2 ERF is roughly 30% smaller than the IRF in the HadGEM2 model.

ERF_nudge values do not reproduce ERF_fSST perfectly. Figure 2 shows that for 2xCO2 and 2%Sol, in particular, global mean ERFs from ERF_nudge are closer to the IRF values, suggesting that the rapid adjustments

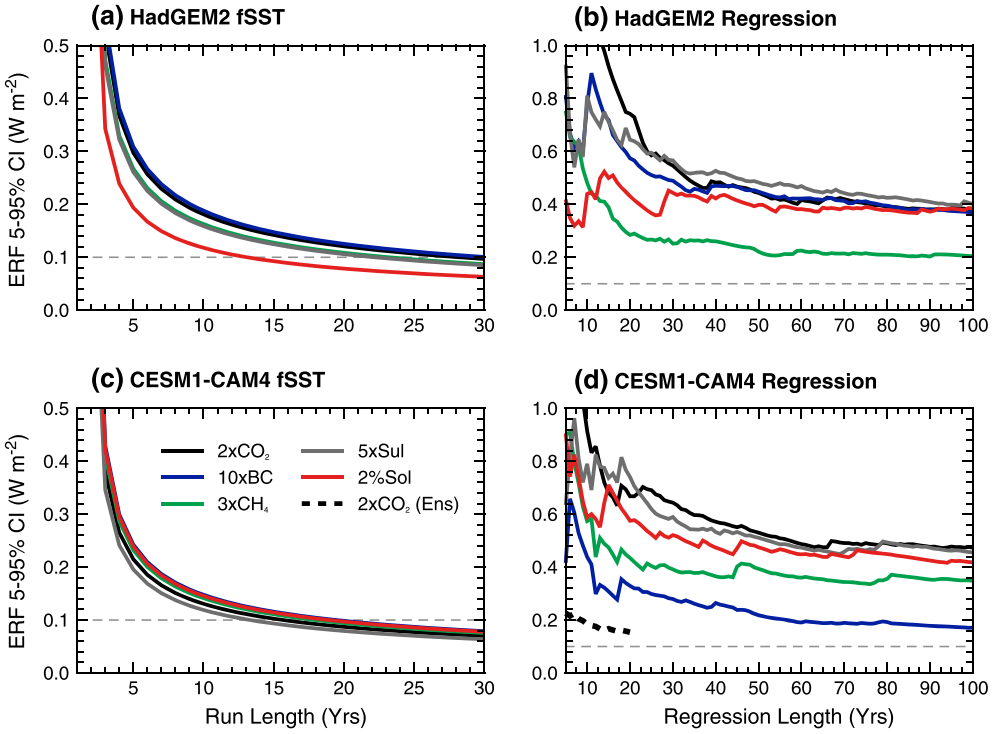

Figure 3. Global (a and c) ERF_fSST and (b and d) ERF_reg 5%–95% confidence interval against run length and regression length, respectively. Results are shown for the different forcings (colored lines) in the two models. A five ensemble member regression out to 20 years is shown as the thick dotted line for *2xCO2* in the CESM1 model.

estimated by the nudging methodology are not as strong as those evaluated by ERF_fSST or ERF_reg. This also suggests that for these forcing mechanisms in this model that the dynamical rapid adjustments that are constrained by ERF_nudge contribute to the total ERF and cannot be neglected (see supporting information for further comparison).

The form of how ERF_fSST uncertainty depends on run length is shown in Figures 3a and 3c. The ERF_fSST uncertainty shows limited variation with forcing mechanism and is largely determined by the variability of the TOA fluxes in the baseline simulation. This variability is slightly larger in HadGEM2 leading to slightly larger uncertainty, but they are more-or-less equivalent in the two models. The 2% solar experiment has noticeably smaller uncertainty in HadGEM2. This appears to be due to reduced variability in this simulation, the cause of which is under investigation. Broadly speaking, 10 years of data could be expected to give global ERF to within 0.2 W m^{-2}, whereas 30 years of data would be needed for 0.1 W m^{-2} accuracy. The accuracy improves with square root of the number of years in the integration (equation (4)). Therefore, diagnosing ERFs of 0.01 W m^{-2} would require centuries of integration and be prohibitive in terms of computer time using this method. Further analysis (not shown) found that variability in TOA flux in the CMIP5 ensemble mean was similar to both models employed here (0.19 W m^{-2}). This indicates that the ERF_fSST uncertainty characteristics of these models are likely representative of the average CMIP5 model. However, two CMIP5 models in our analyses had over 50% higher variability in TOA flux, which indicates that in some models run lengths of considerably over 30 years could be needed to constrain ERF_fSST to within 0.1 W m^{-2}.

Figures 3b and 3d show the uncertainty characteristics of the ERF_reg method. The ERF_reg uncertainty depends mostly on the magnitude of the forcing, except for the 2%Sol experiment in the HadGEM2 model, which has smaller uncertainty. ERF_reg uncertainty is larger than ERF_fSST uncertainty for a single ensemble member and never falls below 0.1 W m^{-2} for the forcings considered here. These uncertainty characteristics can be improved by adding further ensemble members. The dotted line shows a five ensemble member example for 2xCO2 in CESM1 where uncertainties are reduced compared to the single ensemble member. As expected, the reduction in uncertainty scales with the square root of the number of ensemble members. Therefore, to get the ERF_reg 2xCO2 90% confidence interval to less than 0.1 W m^{-2} would likely require

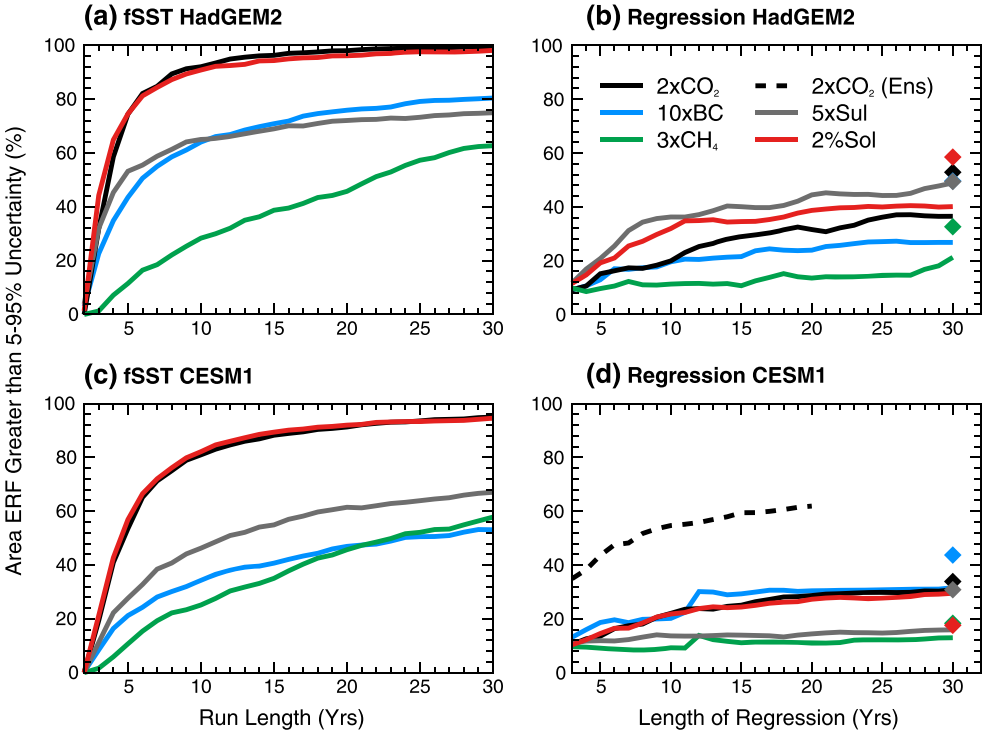

Figure 4. Percentage of the Earth's surface where the local ERF is larger than its local 5%–95% confidence interval against integration length for (left) ERF_fSST and (right) regression length for ERF_reg. Diamonds mark regression values after 100 years. A five ensemble member regression out to 20 years is shown as the thick dotted line for *2xCO2* in the CESM1 model.

more than 10 ensemble members assuming 20 year regressions. This is equivalent to at least 200 years of coupled model integration, a substantial amount of computing resource.

Figure 4 examines the uncertainty in ERF, comparing local ERF to its local 90% confidence interval. The spatial patterns of IRF, ERF, and confidence intervals for ERF_fSST and ERF_reg are presented in the supporting information. For ERF_fSST the robustness of the forcing pattern increases with run length. For run lengths of 30 years ERF_fSST gives an estimate of forcing over more than 50% of the globe that is larger than its confidence interval for the considered experiments. Patterns are not generally as robust from the ERF_reg method, but still have useful skill over about a third of the globe (for 20 year regressions). Adding ensemble members to ERF_reg increases the skill, as expected. The amount of pattern significance obtained increases with the magnitude of the global mean forcing. More spatially, uniform patterns also have higher significance. This explains the lower pattern significance for the 3xCH4 and 10xBC experiments, compared to 2xCO2.

4. Effect of Base Line Climatology and Transient ERF

The utility of the radiative forcing concept relies on the identified roughly linear relationship between forcing and temperature response. Its utility is further strengthened because simple relationships often exist between concentration and forcing. In particular, global mean greenhouse gas forcings can usually be estimated to first order using simple formulae that depend on background concentration levels [*Forster et al.*, 2007], but not on the background climate state [*Ramaswamy et al.*, 2001]. However, different background climate states may give different ERFs for the same trace gas or aerosol perturbation and this is investigated here. The biggest influence on ERF from climatological changes is likely to be due to variations in cloudiness. This may be especially important when trying to diagnose time-varying ERF from the models: as their surface temperatures warm and climates evolve their ERF for a given perturbation might change. This means that the ERF diagnosed with preindustrial SSTs (for example) may not be representative of the ERF experienced at a later date. Conceptually, the ERF that is most representative of a model's response would come from applying

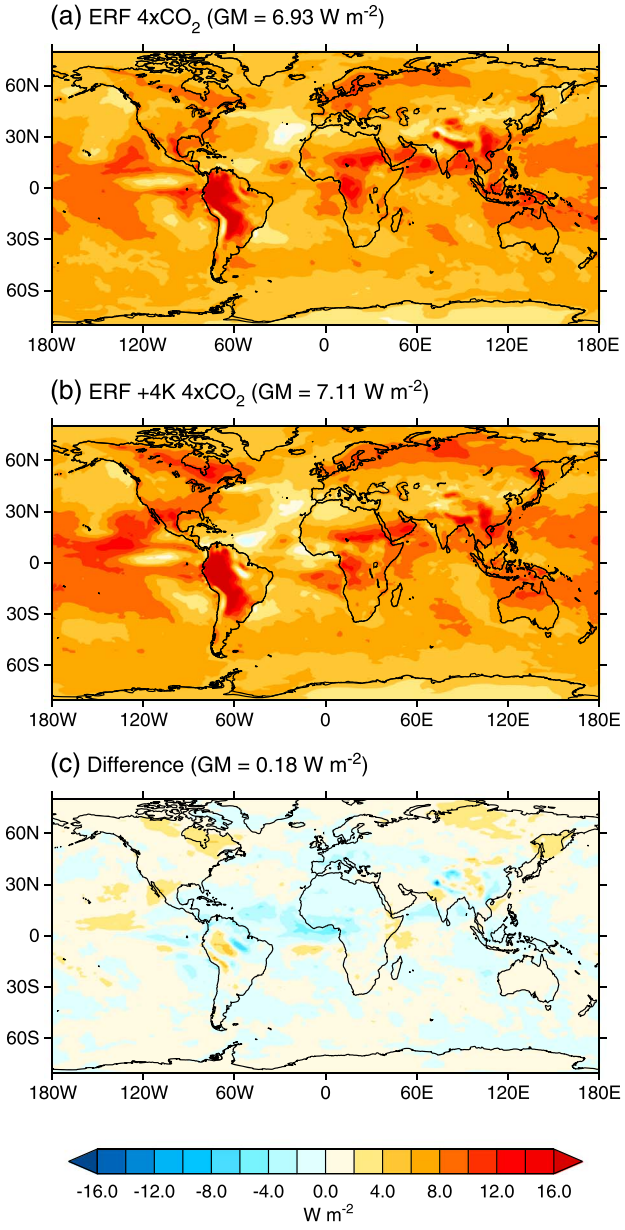

Figure 5. The pattern of 4xCO2 ERF in HadGEM2 employing (a) a preindustrial SST climatology and (b) a preindustrial SST with a uniform 4 K added. (c) The difference between the two is shown. Global mean (GM) values are also shown.

the time-varying SSTs and sea ice from the equivalently forced coupled model run in the ERF calculation. This is the ERF_trans method proposed by AerChemMIP [*Collins et al.*, 2016].

To gauge the possible differences in ERF introduced by different prescriptions of SST and sea ice, we perform two experiments in the HadGEM2 model (Figures 5 and 6) and test the exact AerChemMIP approach in CESM1 (Figure 7).

Figure 5 compares 4xCO2 ERF_fSST generated with a preindustrial climatology of SSTs to ERF_fSST generated with +4 K warmer climatological SSTs. Figure 6 further compares historical ERF_trans for 1979–2008 simulations with sea surface temperature and sea ice taken from a preindustrial climatology (sstClim), the Atmospheric Model Intercomparison Project (AMIP) mean climatology (AMIPClim), and a time-varying monthly AMIP climate state (AMIP).

Figure 5 shows that the pattern of 4xCO2 ERF for the two prescriptions of SST and sea ice is quite similar. Locally (in 1° grid boxes) differences of 20% can be found in ERF over small areas where baseline clouds differ. However, the local effects tend to cancel out in the global mean and the difference in global mean ERF estimates is less than 3%, with the warmer sea surface experiment having a slightly larger globally averaged ERF. The time-varying experiments in Figure 6 do not indicate any systematic differences in ERF_trans caused by employing different SSTs patterns. These results generally indicate that the ERF is relatively insensitive to the prescribed climate of SST and sea ice, if scaled uniformly, even for quite large changes in these fields representative of preindustrial and future conditions.

Figure 6 does raise an important question about how to manage variability in ERF. Historical ERF has a significant variability from the rapid adjustment of clouds, which is a driver of short-term variability within the models. The contribution of such random climate change to variability in ERF can be gauged by comparing individual ensemble members. The contribution of SST and sea ice variation in ERF_trans can be estimated by comparing the thick red line to the thick blue line in Figure 6, although given the small number of ensemble members random noise may still contribute to these ensemble averaged differences. Applying time-varying SSTs and sea ice (red lines) instead of annually repeating boundary conditions (black and blue lines) does not

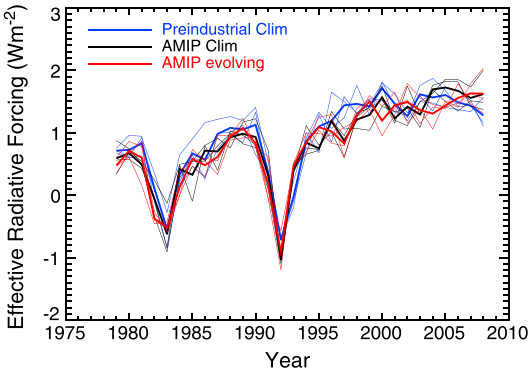

Figure 6. ERF for 1979–2008 estimated by different ERF_trans methods of prescribing SST and sea ice base climates. The base climate states were taken from either a preindustrial climatology (Preindustrial Clim), a 1979–2008 AMIP mean climatology (AMIP Clim), or an evolving monthly AMIP climate (AMIP evolving). ERF was computed from pairs of simulations in HadGEM2, comparing simulations with historical evolution of forcing agents with one employing preindustrial concentrations. The historical simulations include greenhouse gas changes, aerosol, volcanic, and solar forcings but exclude land use changes. Three historical simulations with different atmospheric initial conditions were performed for each base climate state. Thick lines are ensemble averages.

significantly increase the year-to-year variability in ERF, so variation in the SST and sea ice fields appears to be of secondary importance for ERF.

We also tested the ERF_trans method for diagnosing historical sulfate forcing in CESM1, where we employed evolving historical SSTs and sea ice for the base climate state, mimicking the proposed AerChemMIP protocol for computing transient ERF. As described in section 2, ERF_trans was computed as the difference between two integrations: one with a full set of transient forcers and one where all forcers evolved but the atmospheric sulfate aerosol concentration was kept at 1850 values. In addition, we performed two 30 year time slice integrations at 1850 and 2000 conditions to diagnose ERF_fSST at the endpoint more accurately.

Figure 7 shows time series of (a) the global surface temperature anomaly, (b) the global mean sulfate burden, and (c) the resulting ERF_trans. The ERF_fSST for year 2000 calculated from the time slice integrations is shown by the whisker bar in Figure 7c. In addition, Figure 7c shows the ERF estimated assuming a constant forcing per gram calculated separately using the ERF_fSST method with annually repeating preindustrial SSTs and sea ice. For this particular model, interannual variability in ERF_trans has a standard deviation of 0.6 W m^{-2}. Both interannual variation in SSTs and additional random interannual variation in clouds might contribute to this overall variability. The large year-to-year variation clearly poses challenges for diagnosing real short-term changes in ERF. As found for HadGEM2, there appears very little bias between ERF_trans and ERF_fSST (comparing green and black estimates for the present day). Linear scaling of burden appears to work well for estimating the time evolution of ERF (comparing pale blue and black lines).

A main finding is that the year-to-year variability in ERF_trans makes it difficult to isolate possible real departures of ERF trend away from linear scaling, such as that seen between 1880 and 1910. The explanation for

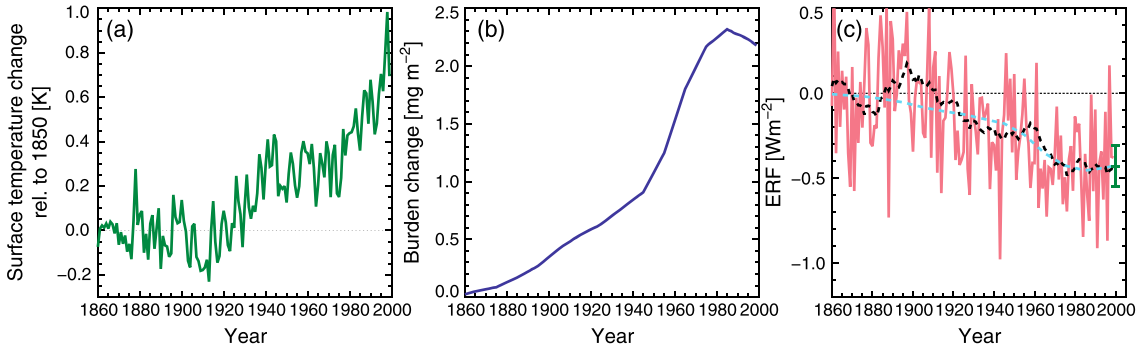

Figure 7. Illustrates sulfate forcing experiments in CESM1 to compute ERF with the ERF_trans method and compare it with ERF_fSST methods. (a) Global mean temperature anomaly from historically evolving prescribed SSTs, (b) sulfate burden, (c) annually averaged ERF_trans (red), and its 15 year running mean (black). The green whisker bar shows the ERF from a 30 year ERF_fSST integration with the year 2000 burden (5%–95% confidence interval). Also shown in Figure 7c is ERF estimated by calculating the global mean ERF per global mean burden for this model derived by the ERF_fSST method (pale blue line).

this behavior is not immediately apparent. Multiple ensemble members may help by reducing the noise to manageable levels.

5. Discussion

Figure 2 shows that there is evidence of some systematic variation between the ERF derived from different approaches. As expected, ERF estimates also generally differ from IRF estimates indicating the importance of rapid adjustments. Previous studies have found that ERF_fSST is systematically larger than ERF_reg for 4xCO2 experiments across a range of models [*Andrews et al*., 2012; *Chung and Soden*, 2015]. Our results show that such a bias is only likely true for long regression lengths (e.g., the 150 year regressions employed in *Andrews et al*. [2012]); short regression lengths give higher ERF_reg estimates. Differences between the methods are particularly apparent for increases in solar constant in both models. Here we discuss the pros and cons of the different methods before making our recommendation. These are summarized in Table 1.

5.1. IRF

Pros This is expected to have less spread across models when measured at the tropopause because it does not depend upon rapid model-dependent feedbacks that contribute to the ERF and is directly comparable with offline estimates of instantaneous radiative forcing, thereby providing a direct test of a model's radiative transfer code. It does not require long model integrations and is the computationally cheapest way to estimate forcing as only radiative transfer codes are required. When used in conjunction with an ERF estimate, it provides a quantification of rapid adjustments. Among the methods discussed in this study, this is the only method for quantifying small forcings ($\lesssim$0.2 W m^{-2}) without a large uncertainty relative to the forcing.

Cons It excludes the effect of rapid adjustments, including stratospheric adjustment and changes to tropospheric clouds and land surface, all of which are known to influence ERF significantly. Special code is also required.

5.2. ERF_fSST Method

Pros ERF_fSST can be diagnosed from short integrations in atmospheric models that do not require ocean coupling. It has the best uncertainty characteristics of methods that include rapid adjustments: a 30 year integration can constrain global forcing to within 0.1 W m^{-2}. Regionality is also well sampled: 2xCO2 forcing has a signal greater than its 90% confidence interval over 90% of the globe. Different SST formulations can be tested, and ERF can be diagnosed before a coupled run is undertaken adding to its use in model and scenario development.

Cons Land surface temperatures change lead to a degree of global mean surface temperature change making the forcing less applicable to a simple global framework that uses global mean temperature to separate forcing and response. Approaches that fit the ERF_fSST into such a framework by making a global averaged feedback based "correction" to the land surface-based global temperatures change do not work well [*Hansen et al*., 2005; *Sherwood et al*., 2015]. A modified fSST approach has also been tried where both land and sea surface temperatures are fixed to climatological values [*Shine et al*., 2003]. However, this cannot readily be implemented in models with sophisticated land surface schemes that need to capture the diurnal cycle in soil temperatures. To quantify the ERF_fSST with 30 year integration with reasonable accuracy, the forcing magnitude should be larger than 0.1 W m^{-2}, otherwise longer simulations are needed.

5.3. ERF_reg Method

Pros Both forcing and response can be diagnosed from a single model integration. Regression to $T = 0$ is also conceptually attractive from a global energy budget point of view, as it separates driver mediated responses (forcings) from global mean surface temperature mediated responses (feedbacks) [*Sherwood et al*., 2015]. As it is computed directly from coupled model integrations, it maybe more akin to the forcing realized within the actual modeling framework employed to gauge climate response.

Cons ERF_reg cannot be readily be determined for time-dependent forcing scenarios as step change experiments requiring many years of model integration would be needed at each time slice considered. ERF_reg also depends on the choice of number of years of data used in the regression analysis, as well as the type of regression. These choices are rather subjective and, as different models would be expected to have different time-dependent feedbacks, it is not clear how the regression length or regression model could be standardized to give a comparable forcing across different models [*Andrews et al*., 2015]. Uncertainties are large for a single ensemble member, and large ensembles would be needed to match the confidence interval of

Journal of Geophysical Research: Atmospheres

10.1002/2016JD025320

Table 1. Characteristics of the IRF, ERF_fSST, ERF_nudge, ERF_reg, and ERF_trans Methods

	IRF	ERF_fSST	ERF_nudge	ERF_reg	ERF_trans
Method	Tropopause flux differences; radiation scheme called twice in online model, once with current concentrations and once with a greenhouse gas or aerosol set to its baseline value	TOA flux differences between two integrations of an atmosphere GCM with interactive land surface but the same prescribed sea ice and sea surface temperatures in the two integrations	TOA flux differences between a perturbation and a control simulation, where both are nudged to meteorology from either a control run of the same model or reanalysis data	Perform a step change forcing experiment in an AOGCM and regress annual TOA flux against T. ERF is intercept at $T = 0$	A variant of ERF_fSST employed in transient runs, ERF is the TOA flux difference between two transient integrations. The same prescribed sea ice and sea surface temperatures are used in the two integrations
Methodological choices	Level to record fluxes and choice of baseline concentrations	The sea ice and SST fields to use for the control state	Whether to include or neglect atmospheric temperature; bottom and top model levels for nudging; relaxation timescale; and control/ observational data set	The control climate state; number of years to include in regression	As for ERF_fSST; in particular choosing whether or not to evolve prescribed SST and sea ice fields
Effect of methodological choices	For drivers that have significant stratospheric adjustment, level of derived forcing is important	Sea ice and SST fields: O(20% effect on regional (1° grid boxes) ERFs; O(3%) on global ERF	Could improve error characteristics by factor of 10 (not evaluated in this study); may introduce systematic bias by not accounting for dynamical feedbacks	ERF generally decreases with the years used in the regression which gives some confounding of forcing and feedback	Using evolving SSTs adds to random error in ERF at a given time but may also capture some real-world variability in ERF
Run lengths for basic global estimate	1 year	5 years of atmosphere model integration times 2	1 year of atmosphere model integration times 2	20 years of coupled model integration times 2	Transient atmosphere model run length times 2 (e.g., 330 years for a 1850–2015 historical run), plus a parallel coupled run if evolving model-specific SSTs are needed
Run lengths required for 90% confidence interval $< 0.1\ W\ m^{-2}$	1 year; there is very little random error	30 years of model integration times 2	6 years (times 2) provide 90% confidence interval $< 0.1\ W\ m^{-2}$ for all experiments except 5xSul	900+ years; for 20 year regressions, achieved via multiple ensembles	Single year ERF estimate has uncertainty of ~$0.6\ W\ m^{-2}$. Can reduce uncertainty by averaging over longer period (36 years) or having 36 ensembles
Percentage of Earth with characterized forcing for 20 year integrations	100%	90% + with ERF larger than the 90% confidence interval for 2xCO2 forcing	Not estimated: likely to be close to 100%	25% + with ERF larger than the local 90% confidence interval for 2xCO2 forcing	Not estimated
Limitations	By design it excludes rapid adjustments; requires explicitly coding into climate models	A large number model simulations needed to gauge ERF time evolution if making multiple time slices	Requires sophisticated code. Subjective methodological choices can have large impact on results; potential bias from exclusion of dynamical adjustments	Computationally prohibitive for ERF time evolution quantification	Large random errors; hard to distinguish real-world year to year change from noise

ERF_fSST. Regional uncertainties would be larger still. The computer resources needed for such large ensembles would be prohibitively large, making ERF_reg unsuitable for small forcings or distinguishing small forcing differences.

5.4. ERF_nudge Method

*Pros*As with ERF_fSST, only a few years of model integration are required to obtain a good estimate of ERF. In addition, interannual variability in TOA flux differences is much reduced compared to ERF_fSST: a variant of ERF_fSST from a 10 year nudged experiment has only around one tenth of the standard error from a free-running integration [*Kooperman et al*., 2012, Figure 2d].

Cons ERFs estimated from nudging likely fail to account for any rapid adjustment in the circulation. Such nudged ERFs are therefore not strictly equivalent to ERF_fSST, and it is an open question as to whether this could lead to significant global and/or regional systematic biases between nudged and standard methods. Additionally, implementing nudging is complex and requires a modification to climate model code. Nudging also requires subjective choices in relation to which model levels and variables should be nudged, as well as the choice of model/reanalysis data set to nudge to.

5.5. ERF_trans Method

Pros Can provide transient ERF estimates from paired integrations without the need for multiple time slice integrations. Paired integrations, which are performed for other reasons, can be checked easily for forcing and feedbacks. By using time evolving SSTs, it has the potential to account for some SST nonlinear effects, such as ERF dependence on cloud base state. Decadal changes in ERF may be detected if the forcing signal is large enough.

Cons The year-to-year random error is expected to be large requiring multiple ensemble members and/or decadal averaging to detect typical forcing trends. Any effect of SST base state on forcing is likely to be small and swamped by random error. This makes the use of time evolving SSTs unhelpful for determining year-to-year changes in ERF without a large number of ensembles.

6. Summary and Recommendations

If our ambition is to estimate ERF to within 0.1 W m^{-2} globally, the ERF_fSST approach meets this goal with 30 years of model integration. To obtain similar uncertainty characteristics from the ERF_reg approach would require more than 45 ensemble members of 20 year coupled model integrations, making ERF_fSST a far more computationally efficient method.

The attractiveness of the ERF_reg method lies in its simple split of forcing and response, whereby ERF_reg does not have a contribution from a global mean temperature change. ERF_reg is also calculated within the same coupled modeling framework used to model the response so may be more representative of the energy budget changes felt within coupled model integrations. These potential conceptual benefits are limited, though, as simple global linear regression is unlikely to capture the intricacies and multiple timescale of model response. Different regression approaches could be used, but there is no clear justification for choosing between methods. There are also conceptual issues with the fSST approach as this method incorporates a residual global temperature change that does not fit a simple global temperature driven feedback model. The conceptual issues with both ERF_reg and ERF_fSST really show the limitations of a forcing/response framework based on the global energy budget. In reality processes are not divided between forcing and response and they affect both TOA energy budget and surface temperature on many different timescales. Nevertheless, we argue that ERF, however it is calculated, presents a useful first-order comparison of how climate models "feel" an anthropogenic or natural perturbation.

ERF_fSST has a clear advantage over ERF_reg in terms of the computational overhead needed to reduce uncertainties to workable levels. This computational overhead makes ERF_fSST approaches far more suitable for quantifying forcings across drivers and/or models and examining regional differences in forcing. ERF_trans is a variation of the ERF_fSST method and can be usefully employed to quantify time-varying ERFs.

Chung and Soden [2015] explored three types of ERF_fSST estimates for 4xCO2 within CMIP5 models. The approaches used different base states based on the model's climatological SSTs, an AMIP-based climatology, or aqua planet simulations. There were differences in the computed ERF_fSST between the base states but no strong bias, consistent with our results from section 4. Using annually repeating climatological SSTs and sea

ice as the base state has the main advantage that they are periodic, so the control experiments need only be run for 30 years, rather than 240 years to cover 1860 to 2100 scenarios. They also have the advantage of being easier/smaller for modeling groups to produce as boundary conditions for their atmospheric models. Employing an AMIP-based SST and sea ice climatology as in *Andrews* [2014] or *Chung and Soden* [2015] has the advantage of being able to piggyback off exiting AMIP runs that are routinely performed by modeling centers. The ERF may also be more representative of time period being analyzed. However, if a coupled simulation had different SSTs than the AMIP simulation, an AMIP-derived ERF would not represent the forcing within an individual model. Further, as AMIP simulations are based on observed SSTs, it is not straightforward to adapt the method for future forcings. Ideally, one might use an ERF_trans method where SSTs are employed from the partner coupled model integration, either individual ensemble members or the ensemble average (as proposed in AerChemMIP). This means that one may need to perform long coupled model integrations before the simulations to diagnose forcing can be done, limiting the use of forcing diagnostics as a model development tool. There would only be an incentive to add this degree of complication if there was clear evidence that there was a strong forcing dependence on SST or if one wanted specifically to capture short-term variability in ERF. Our tests of ERF_trans show that any global SST dependent effect is likely small and swamped by year-to-year random error, although, there may be greater regional differences for individual forcings, e.g., the BC on snow forcing would depend on snow cover in a given climate. On the other hand, if there were other reasons to perform ERF_trans-like simulations, decadal averaged forcing may be derived from these simulations.

Generally, ERF_fSST and ERF_trans were very insensitive to the choice of SST and sea ice base climate. This is encouraging for comparing ERF in studies that employ different SST and sea ice base climate formulations. A particular example arises when comparing ERF derived from concentration and emission-based models, as concentration based models routinely use preindustrial base states and emission-based models often use present-day base states to take advantage of being able to use observed wind fields.

We recommend adopting an ERF_fSST approach for ERF estimates going forward, using a fixed SST and sea ice seasonally varying climatology based on the model's own preindustrial climatology. This method is currently proposed by RFMIP for both its time slice simulations and for historic and future scenarios. AerChemMIP also currently proposes a fixed preindustrial climatology for its time slice experiments. The free-running integrations are suitable for diagnosing ERF to within a 5% to 95% confidence interval of 0.1 W m^{-2} for a 30 year simulation. If better accuracy is needed, nudging approaches may provide a useful way forward but are hard to implement and have not yet been sufficiently tested in a range of models.

IRFs are a poor representation of ERF but are nevertheless useful to compute. First, they provide a set of very useful tests of a model's radiative transfer code, allowing a more direct comparison with other models and also with sophisticated line by line radiative transfer codes [*Collins et al*., 2006; *Forster et al*., 2011; *Oreopoulos et al*., 2012; *Pincus et al*., 2015]. Second, computing both IRF and ERF within the same model allows an accurate quantification of the effects of rapid adjustment. The standard IRF method employs a second radiation call under preindustrial concentrations to give an estimate of direct forcing that be compared with ERF to estimate rapid adjustment. Additional radiation calls with aerosol scattering and absorption neglected, and cloud scattering and absorption neglected, to distinguish the forcing from aerosol-cloud interaction from aerosol-radiation interaction are needed to distinguish types of rapid adjustment [*Ghan*, 2013; *Ghan et al*., 2012].

AerChemMIP will use time-varying SSTs from one coupled model ensemble member in its ERF_trans experiments in an effort to achieve a better representation of the evolution of ERF changes realized in the coupled model integration. The approach allows the study of changes in the chemical composition with cheaper model integrations omitting the ocean component. Our results show that random interannual variability will likely swamp any signal of SST-driven interannual forcing change. Our results also show that using a simpler approach of a standard fixed SST and sea ice climatology (as employed in RFMIP) is unlikely to create a systematic bias. We also give a strong cautionary note over the signal to noise ratio that it is possible to achieve with any ERF_trans method. One ensemble member can only pick up large interannual changes in forcing ($>$0.6 W m^{-2}), such as that from volcanic eruptions (Table 1). Three ensemble members would be able to detect globally averaged decadal variations in forcing larger than 0.1 W m^{-2} per decade (Figure 3 and Table 1).

We recommend that both RFMIP and AerChemMIP carefully consider the expected noise characteristics from their ERF_fSST 30 year integrations and plan their analyses accordingly. For ERF_trans, RFMIP are proposing four transient experiments of three ensembles each with (i) all anthropogenic and natural changes, (ii) greenhouse gas changes, (iii) natural changes, and (iv) aerosol changes. The ERFs in these experiments are expected to differ by more than the detectable limit of 0.1 W m^{-2} per decade, so their analyses should prove useful. AerChemMIP are currently proposing further simulations to diagnose ERF_trans due to changes in methane, nitrous oxide, tropospheric ozone precursors, and stratospheric ozone depleting substances. We note that these simulations have other purposes as well, such as a characterization of atmospheric composition in a changing climate. With respect to forcing characterization, it is questionable as to whether the proposed ERF_trans method would be able to detect decadal forcing changes from such experiments given the magnitude of the random errors highlighted here. We therefore suggest that AerChemMIP includes IRF estimates for small forcings and/or considers long ERF_fSST time slice integrations where rapid adjustments are deemed important.

Our paper has clearly shown the value in carefully testing proposed model comparison protocols. We have tested some of the proposed RFMIP and AerChemMIP protocols. We urge other proposed CMIP6 MIPs to carefully test methods and proposed analyses before making firm requests to the world's climate modeling community. For ERF, we recommend adopting a simple and parsimonious approach based on ERF_fSST integrations. This will encourage many more modeling groups to routinely estimate ERF from their simulations.

Acknowledgments

The authors thank Ron Miller, Editor Steve Ghan, and two anonymous referees for thorough and extremely helpful comments that greatly improved the manuscript. P.M.F., C.J.S., and R.P. received funding from Regional and Global Climate Modeling Program of the U.S. Department of Energy Office of Environmental and Biological Sciences under grant DE-SC0012549. P.M.F. and A.C.M. received funding from NERC grant NE/N006038/1 and P.M.F. received additional support from a Royal Society Wolfson Merit Award. T.R. was supported by NERC CASE award NE/K007483/1. G.M. and B.H.S. received funding from the Norwegian Research Council project NAPEX (project 229778). G.M. and M.S. benefitted from the Norwegian research council projects 235548 (Role of SLCF in Global Climate Regime) and 229796 (AeroCom-P3). T.A. was supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101). Data are available on the PDRMIP website data access page http://www.cicero.uio.no/en/PDRMIP/PDRMIP-data-access.

References

Andrews, T. (2014), Using an AGCM to diagnose historical effective radiative forcing and mechanisms of recent decadal climate change, *J. Clim.*, *27*(3), 1193–1209, doi:10.1175/JCLI-D-13-00336.1.

Andrews, T., and M. A. Ringer (2014), Cloud feedbacks, rapid adjustments, and the forcing–response relationship in a transient CO_2 reversibility scenario, *J. Clim.*, *27*(4), 1799–1818, doi:10.1175/JCLI-D-13-00421.1.

Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012), Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, *Geophys. Res. Lett.*, *39*, L09712, doi:10.1029/2012GL051607.

Andrews, T., J. M. Gregory, and M. J. Webb (2015), The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models, *J. Clim.*, *28*(4), 1630–1648, doi:10.1175/JCLI-D-14-00545.1.

Armour, K. C., C. M. Bitz, and G. H. Roe (2012), Time-varying climate sensitivity from regional feedbacks, *J. Clim.*, *26*(13), 4518–4534, doi:10.1175/JCLI-D-12-00544.1.

Boucher, O., et al. (2013), Clouds and Aerosols, 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*, edited by T. F. Stocker, et al., pp. 571–658, Cambridge Univ. Press, Cambridge, and New York, doi:10.1017/CBO9781107415324.016.

Chung, E.-S., and B. J. Soden (2015), An assessment of methods for computing radiative forcing in climate models, *Environ. Res. Lett.*, *10*(7), 074,004.

Collins, M., et al. (2013), Long-term climate change: Projections, commitments and irreversibility, 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*, edited by T. F. Stocker, et al., pp. 1029–1136, Cambridge Univ. Press, Cambridge, and New York, doi:10.1017/CBO9781107415324.024.

Collins, W. D., et al. (2006), Radiative forcing by well-mixed greenhouse gases: Estimates from climate models in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4), *J. Geophys. Res.*, *111*, D14317, doi:10.1029/2005JD006713.

Collins, W. J., et al. (2016), AerChemMIP: Quantifying the effects of aerosols and chemistry in CMIP6, *Geosci. Model Dev. Discuss.*, doi:10.5194/gmd-2016-139, *in review*.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*(656), 553–597, doi:10.1002/qj.828.

Forster, P. M., and K. E. Taylor (2006), Climate forcings and climate sensitivities diagnosed from coupled climate model integrations, *J. Clim.*, *19*(23), 6181–6194, doi:10.1175/jcli3974.1.

Forster, P. M., et al. (2007), Changes in atmospheric constituents and in radiative forcing, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by D. Qin et al., Cambridge Univ. Press, Cambridge and New York.

Forster, P. M., et al. (2011), Evaluation of radiation scheme performance within chemistry climate models, *J. Geophys. Res.*, *116*, D10302, doi:10.1029/2010JD015361.

Forster, P. M., T. Andrews, P. Good, J. M. Gregory, L. S. Jackson, and M. Zelinka (2013), Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models, *J. Geophys. Res. Atmos.*, *118*, 1139–1150, doi:10.1002/jgrd.50174.

Fyfe, J. C., et al. (2016), Making sense of the early-2000s warming slowdown, *Nat. Clim. Change*, *6*(3), 224–228, doi:10.1038/nclimate2938.

Ghan, S. J. (2013), Technical note: Estimating aerosol effects on cloud radiative forcing, *Atmos. Chem. Phys.*, *13*(19), 9971–9974.

Ghan, S. J., X. Liu, R. C. Easter, R. Zaveri, P. J. Rasch, J.-H. Yoon, and B. Eaton (2012), Toward a minimal representation of aerosols in climate models: Comparative decomposition of aerosol direct, semidirect, and indirect radiative forcing, *J. Clim.*, *25*(19), 6461–6476, doi:10.1175/JCLI-D-11-00650.1.

Gregory, J. M., and T. Andrews (2016), Variation in climate sensitivity and feedback parameters during the historical period, *Geophys. Res. Lett.*, *43*, 3911–3920, doi:10.1002/2016GL068406.

Gregory, J. M., and P. M. Forster (2008), Transient climate response estimated from radiative forcing and observed temperature change, *J. Geophys. Res.*, *113*, D23105, doi:10.1029/2008JD010405.

Gregory, J. M., W. J. Ingram, M. A. Palmer, G. S. Jones, P. A. Stott, R. B. Thorpe, J. A. Lowe, T. C. Johns, and K. D. Williams (2004), A new method for diagnosing radiative forcing and climate sensitivity, *Geophys. Res. Lett.*, *31*, L03205, doi:10.1029/2003GL018747.

Hansen, J., et al. (2005), Efficacy of climate forcings, *J. Geophys. Res.*, *110*, D18104, doi:10.1029/2005JD005776.
IPCC (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1535 pp., Cambridge Univ. Press, Cambridge and New York, doi:10.1017/CBO9781107415324.
Kooperman, G. J., M. S. Pritchard, S. J. Ghan, M. Wang, R. C. J. Somerville, and L. M. Russell (2012), Constraining the influence of natural variability to improve estimates of global aerosol indirect effects in a nudged version of the Community Atmosphere Model 5, *J. Geophys. Res.*, *117*, D23204, doi:10.1029/2012JD018588.
Larson, E. J. L., and R. W. Portmann (2016), A temporal kernel method to compute effective radiative forcing in CMIP5 transient simulations, *J. Clim.*, *29*(4), 1497–1509, doi:10.1175/JCLI-D-15-0577.1.
Marotzke, J., and P. M. Forster (2015), Forcing, feedback and internal variability in global temperature trends, *Nature*, *517*(7536), 565–570, doi:10.1038/nature14117.
Marvel, K., G. A. Schmidt, R. L. Miller, and L. S. Nazarenko (2015), Implications for climate sensitivity from the response to individual forcings, *Nat. Clim. Change*, *6*, 386–389, doi:10.1038/nclimate2888.
Meraner, K., T. Mauritsen, and A. Voigt (2013), Robust increase in equilibrium climate sensitivity under global warming, *Geophys. Res. Lett.*, *40*, 5944–5948, doi:10.1002/2013GL058118.
Myhre, G., and F. Stordal (1997), Role of spatial and temporal variations in the computation of radiative forcing and GWP, *J. Geophys. Res.*, *102*, 11,181–11,200, doi:10.1029/97JD00148.
Myhre, G., et al. (2013), Anthropogenic and natural radiative forcing, 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*, edited by T. F. Stocker, et al., pp. 659–740, Cambridge Univ. Press, Cambridge and New York, doi:10.1017/CBO9781107415324.018.
Oreopoulos, L., et al. (2012), The continual intercomparison of radiation codes: Results from phase I, *J. Geophys. Res.*, *117*, D06118, doi:10.1029/2011JD016821.
Pincus, R., et al. (2015), Radiative flux and forcing parameterization error in aerosol-free clear skies, *Geophys. Res. Lett.*, *42*, 5485–5492, doi:10.1002/2015GL064291.
Pincus, R., P. M. Forster, and B. Stevens (2016), The Radiative Forcing Model Intercomparison Project (RFMIP): Experimental protocol for CMIP6, *Geosci. Model Dev. Discuss.*, doi:10.5194/gmd-2016-88, *in review*.
Ramaswamy, V., O. Boucher, J. Haigh, D. Hauglustaine, J. Haywood, G. Myhre, T. Nakajima, G. Y. Shi, and S. Solomon (2001), Radiative forcing of climate change, in *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, edited by Y. Ding et al., Cambridge Univ. Press, Cambridge and New York.
Rose, B. E. J., K. C. Armour, D. S. Battisti, N. Feldl, and D. D. B. Koll (2014), The dependence of transient climate sensitivity and radiative feedbacks on the spatial pattern of ocean heat uptake, *Geophys. Res. Lett.*, *41*, 1071–1078, doi:10.1002/2013GL058955.
Rugenstein, M. A. A., J. M. Gregory, N. Schaller, J. Sedláček, and R. Knutti (2016), Multiannual ocean–atmosphere adjustments to radiative forcing, *J. Clim.*, *29*(15), 5643–5659, doi:10.1175/JCLI-D-16-0312.1.
Samset, B. H., et al. (2013), Black carbon vertical profiles strongly affect its radiative forcing uncertainty, *Atmos. Chem. Phys.*, *13*(5), 2423–2434, doi:10.5194/acp-13-2423-2013.
Sherwood, S., P. M. Forster, J. Gregory, S. Bony, B. Stevens, and C. Bretherton (2015), Adjustments in the forcing feedback framework for understanding climate change, *Bull. Am. Meteorol. Soc.*, *96*, 217–228, doi:10.1175/BAMS-D-13-00167.1.
Shindell, D. T. (2014), Inhomogeneous forcing and transient climate sensitivity, *Nat. Clim. Change*, *4*(4), 274–277, doi:10.1038/nclimate2136.
Shine, K. P., J. Cook, E. J. Highwood, and M. M. Joshi (2003), An alternative to radiative forcing for estimating the relative importance of climate change mechanisms, *Geophys. Res. Lett.*, *30*(20), 2047, doi:10.1029/2003GL018141.
Stevens, B. (2015), Rethinking the lower bound on aerosol radiative forcing, *J. Clim.*, *28*(12), 4794–4819, doi:10.1175/JCLI-D-14-00656.1.
Storelvmo, T., T. Leirvik, U. Lohmann, P. C. B. Phillips, and M. Wild (2016), Disentangling greenhouse warming and aerosol cooling to reveal Earth's climate sensitivity, *Nat. Geosci.*, *9*, 286–289, doi:10.1038/ngeo2670.
Telford, P. J., P. Braesicke, O. Morgenstern, and J. A. Pyle (2008), Technical note: Description and assessment of a nudged version of the new dynamics unified model, *Atmos. Chem. Phys.*, *8*, 1701–1712, doi:10.5194/acp-8-1701-2008.
Vial, J., J. L. Dufresne, and S. Bony (2013), On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates, *Clim. Dynam.*, *41*(11-12), 3339–3362.
Zelinka, M. D., T. Andrews, P. M. Forster, and K. E. Taylor (2014), Quantifying components of aerosol-cloud-radiation interactions in climate models, *J. Geophys. Res. Atmos.*, *119*, 7599–7615, doi:10.1002/2014jd021710.