Evaluating climate models' cloud feedbacks against expert judgement

Mark D. Zelinka[1], **Stephen A. Klein**[1], **Yi Qin**[1]

[1]Lawrence Livermore National Laboratory

Key Points:

- Models with smallest feedback errors have moderate total cloud feedbacks and ECS
- Models with large positive total cloud feedbacks have several systematically high-biased components
- Better simulation of mean-state cloud properties leads to stronger but not necessarily better cloud feedbacks

Corresponding author: Mark D. Zelinka, zelinka1@llnl.gov

Abstract

The persistent and growing spread in effective climate sensitivity (ECS) across global climate models necessitates rigorous evaluation of their cloud feedbacks. Here we evaluate several cloud feedback components simulated in 19 climate models against benchmark values determined via an expert synthesis of observational, theoretical, and high-resolution modeling studies. We find that models with smallest feedback errors relative to these benchmark values have moderate total cloud feedbacks (0.4–0.6 $\mathrm{Wm^{-2}K^{-1}}$) and generally moderate ECS (3–4 K). Those with largest errors generally have total cloud feedback and ECS values that are too large or too small. Models tend to achieve large positive total cloud feedbacks by having several cloud feedback components that are systematically biased high rather than by having a single anomalously large component, and vice versa. In general, better simulation of mean-state cloud properties leads to stronger but not necessarily better cloud feedbacks. The Python code base provided herein could be applied to developmental versions of models to assess cloud feedbacks and cloud errors and place them in the context of other models and of expert judgement in real-time during model development.

Plain Language Summary

Climate models strongly disagree with each other regarding how much warming will occur in response to increased greenhouse gases in the atmosphere. This is mainly because they disagree on the response of clouds to a warming planet — a process known as the cloud feedback that can amplify or dampen warming initially caused by carbon dioxide. In this study we compare many models' cloud feedbacks to those that have been determined by a recent expert assessment of the literature. We find that the models whose cloud feedbacks most strongly disagree with expert assessment tend to have more extreme cloud feedbacks and hence warm too much or too little in response to carbon dioxide. The models with total cloud feedbacks that are too large do not have a single massive feedback component but rather several components that are larger than in other models. Models that simulate current-climate clouds that look more like those in nature also simulate stronger amplifying cloud feedbacks, but doing a better job at simulating current-climate clouds does not, in general, lead to a better simulation of cloud feedbacks.

1 Introduction

Cloud feedback — the change in cloud-induced top-of-atmosphere radiation anomalies with global warming — is the primary driver of differences in effective climate sensitivity (ECS) across global climate models (GCMs). This has been the case for all existing model intercomparisons, starting with Cess et al. (1989); Cess and others (1990) and continuing to the most recent collection of models as part of CMIP6, the 6th phase of the Coupled Model Intercomparison Project (Zelinka et al., 2020; Eyring et al., 2016). Despite substantial progress in understanding, diagnosing, modeling, and observationally constraining cloud feedbacks from a variety of approaches, the spread in cloud feedbacks across GCMs has remained substantial through the decades and actually increased in CMIP6 relative to CMIP5 (Zelinka et al., 2020). Moreover, strengthened cloud feedback — particularly for extratropical low clouds — is the primary reason for the increase in average climate sensitivity in CMIP6 relative to CMIP5, as well as for the emergence of models with very high ECS above the upper limit of the *likely* range (1.5–4.5 K) reported in the fifth assessment report of the Intergovernmental Panel on Climate Change (Collins et al., 2013).

This motivates a desire to evaluate models' cloud feedbacks against some form of ground truth. Such an evaluation is now possible because quantitative values of individual cloud feedbacks (and their uncertainties) were recently determined based on an ex-

pert synthesis of theoretical, observational, and high-resolution cloud modeling evidence. This synthesis was conducted as part of a broader assessment of climate sensitivity, in which three semi-independent lines of evidence (process studies, historical climate record, and paleoclimate record) were brought together in a Bayesian framework to place robust bounds on Earth's climate sensitivity (Sherwood et al., 2020).

Our goals in this work are several-fold. First, we evaluate GCM cloud feedback components against those assessed in Sherwood et al. (2020). This allows us to answer several questions, including: Do models with extremely large or small climate sensitivities have cloud feedback components that are erroneous? If so, which component(s)? How are cloud feedbacks in CMIP6 — and their biases with respect to expert assessment — changing from CMIP5? Are some models getting the "right" total cloud feedback via erroneous components that compensate?

Second, we investigate whether the fidelity with which models simulate present-day cloud properties is linked to their cloud feedbacks and to the fidelity with which their cloud feedbacks agree with expert judgement. A key question is whether better simulation of present-day cloud properties leads to cloud feedbacks that are better aligned with expert judgement. This is particularly relevant because aspects of the cloud simulation in many high-ECS CMIP6 models are in many cases considered superior to those in CMIP5 (Gettelman et al., 2019; Bodas-Salcedo et al., 2019), yet holistic aspects of the climate simulation in these models appear inferior to their lower-ECS counterparts (Zhu et al., 2020, 2021; Tokarska et al., 2020; Nijsse et al., 2020)

Finally, we provide a code base to compute cloud feedbacks and error metrics for all of the assessed categories, and visualize them in a multi-model context. This will allow, for example, model developers to evaluate cloud feedbacks in developmental versions of their models against expert judgement, other models, and other variants of their model, providing them with detailed information about a key process affecting their model's climate sensitivity.

2 Data and Methods

We are primarily interested in cloud feedbacks in response to CO_2-induced global warming, so we make use of abrupt CO_2 quadrupling experiments conducted with fully-coupled GCMs in CMIP5 and CMIP6 (`abrupt-4xCO2`). We first compute cloud radiative anomalies at the top-of-atmosphere (TOA) by multiplying cloud fraction anomalies with cloud radiative kernels (Zelinka et al., 2012a, 2012b). The cloud fraction anomalies needed for this calculation are reported in a matrix of 7 cloud top pressure (CTP) categories by 7 visible optical depth (τ) categories matching the categorization of the International Satellite Cloud Climatology Project (ISCCP; Rossow & Schiffer, 1999). These matrices are produced by the ISCCP simulator (Klein & Jakob, 1999; M. Webb et al., 2001), referred to as `clisccp` in CMIP parlance. Cloud radiative kernels quantify the sensitivity of top-of-atmosphere radiative fluxes to small cloud fraction perturbations in each of these 49 cloud types. Hence the product of the two yields the radiation anomaly from each cloud type, which can be summed over the entire matrix to provide the total cloud radiative anomalies at a given location. Because of the reliance on `clisccp`, we are limited in this study to those models that have successfully implemented the Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulator Package (COSP; Bodas-Salcedo et al., 2011).

Anomalies are computed with respect to the contemporaneous pre-industrial control (`piControl`) simulation, with three exceptions: CNRM-CM6-1, CNRM-ESM2-1, and IPSL-CM6A-LR-INCA did not archive `clisccp` from the `piControl` simulation, so we take this field from `piClim-control`, a 30-year long atmosphere-only simulation that uses

sea-surface temperatures (SSTs) and sea ice concentrations fixed at the model-specific `piControl` climatology (Pincus et al., 2016).

We compute cloud feedbacks by regressing annual mean cloud-radiative anomalies on annual and global mean surface air temperature anomalies over the duration of the 150-year `abrupt-4xCO2` experiment containing all necessary data. In CMIP6, `clisccp` output is available throughout the full duration of the run, whereas in CMIP5 it is typically only available for two non-contiguous 20-year periods, one at the beginning and one at the end of the run.

We focus in this study on feeedbacks estimated from `abrupt-4xCO2` experiments so as to stay consistent with Sherwood et al. (2020), but have repeated all calculations using AMIP experiments with imposed +4K SST perturbations that are spatially uniform (`amip-p4K`) and patterned (`amip-future4K`), as described in the CFMIP protocol (M. J. Webb et al., 2017). Feedbacks in these simulations were computed as cloud radiation anomalies normalized by global mean surface air temperature anomalies between the +4K experiments and the control `amip` experiment. All basic conclusions reported in this study are insensitive to whether we consider feedbacks diagnosed in `amip-p4K`, `amip-future4K`, or `abrupt-4xCO2` experiments.

To distinguish feedbacks occurring in regions of large-scale ascent from those occurring in regions of large-scale descent over tropical oceans, we aggregate all monthly control and perturbed climate fields over the tropical oceans into 10-hPa wide bins of 500 hPa vertical pressure velocity (ω_{500}) following Bony et al. (2004). Anomalies between perturbed and control climates are then performed in ω_{500} space rather than geographic space when computing feedbacks. The resulting feedbacks can be further broken down into dynamic, thermodynamic, and covariance terms (see Bony et al., 2004), but for the purposes of this study, we will consider only their sum, and will further aggregate these to "ascent regions" where $\omega_{500} < 0$ and "descent regions" where $\omega_{500} \geq 0$.

Following Zelinka et al. (2016), we separately quantify feedbacks arising from low, boundary layer clouds and from non-low, free tropospheric clouds, hereafter referred to as "low" and "high" cloud feedbacks, respectively. This is done by performing the cloud feedback calculations using only restricted parts of the `clisccp` histogram: CTPs $>$ 680 hPa for low clouds and CTPs $\leq$ 680 hPa for high clouds. Within these subsets, the cloud feedback is further broken down into (1) the "amount" component due to change in total cloud fraction holding CTP and τ distribution fixed; (2) the "altitude" component due to the change in CTP distribution holding total fraction and τ distribution fixed; and (3) the "optical depth" component due to the change in τ distribution holding the total fraction and CTP distribution fixed (Zelinka et al., 2013, 2016).

In Table 1, we list the central value and 1-σ uncertainty of the cloud feedback components assessed in Sherwood et al. (2020) and describe how we compute them in GCMs in this study. A large amount of observational evidence, based mainly on inter-annual variability, was used to provide quantitative values for the assessed total cloud feedback and several of its individual components. In addition, process-resolving models in the form of large eddy simulations were a key piece of evidence for the strength of tropical marine low cloud feedback, while guidance from theoretical understanding underlies the assessed high cloud altitude, tropical anvil, and land-cloud amount feedbacks. While many of the expert assessed cloud feedbacks are independent of any GCM results, the assessed central value and uncertainty for the high cloud altitude, land cloud amount, and middle latitude marine low cloud amount feedbacks were derived at least partially from GCMs, albeit a collection that included pre-CMIP5 models that are excluded here and that excluded some recently-published CMIP6 models that are included here. Comparing GCM results to these values can therefore be thought of as a quick and economical way of evaluating model feedbacks against the very wide body of evidence that forms the basis of the expert-assessed cloud feedbacks.

Values of effective climate sensitivity (ECS) are taken from Zelinka et al. (2020), updated to include recently-available models. These ECS values are computed in a manner consistent with the cloud feedbacks, by regressing global and annual mean TOA net radiative flux anomalies on global and annual mean surface air temperature anomalies. Anomalies are computed with respect to the contemporaneous `piControl` simulation, except in IPSL-CM6A-LR-INCA, for which we use `piClim-control` because no `piControl` fields are available.

Finally, we compute mean-state cloud error metrics defined in Klein et al. (2013). Briefly, these scalars quantify the spatio-temporal error of several modeled climatological cloud properties with respect to the ISCCP observational climatology: (1) the total cloud fraction for clouds with τ greater than 1.3; (2) the joint distribution of cloud fraction as a function of cloud top pressure and τ for clouds with $\tau > 3.6$; and (3) the radiatively-relevant cloud property errors. The radiatively-relevant errors are computed as for (2), but cloud fraction anomalies with respect to ISCCP are first multiplied by LW, SW, and net (LW+SW) cloud radiative kernels before integrating over month and space. We compute the model climatological cloud properties using Atmospheric Model Intercomparison Project (`amip`) simulations and the observational climatology using the ISCCP HGG product (Young et al., 2018). Both model and observed climatologies are computed over the 26-year period January 1983 to December 2008, when all model simulations and observations overlap, but error metrics are very insensitive to the time period considered. All error metrics are computed between 60°S and 60°N.

3 Results

3.1 GCM Cloud Feedbacks Evaluated Against Expert-Assessed Values

In Figure 1, cloud feedbacks from 7 CMIP5 and 12 CMIP6 models are compared with the assessed values for feedback categories listed above. Each feedback value is scaled by the fractional area of the globe occupied by that cloud type such that summing all components yields the global mean feedback. Each marker is color-coded by its ECS, with the color boundaries corresponding to the 5th, 17th, 83rd, and 95th percentiles of the Baseline posterior PDF of ECS from Table 10 of Sherwood et al. (2020). In Table 2, we list the GCM values and highlight any values that lie outside of the *very likely* (90%) and *likely* (66%) confidence intervals of expert judgement with double and single asterisks, respectively. Supplementary Figures 1-19 are identical to Figure 1, but with individual models highlighted in each figure for better discrimination.

Many models fall within the *likely* range assessed for the high cloud altitude feedback and the multi-model mean is very close to the central assessed value. However, some models have weak high cloud altitude feedbacks that lie below the lower bound of the *likely* (MRI-CGCM3 and MIROC6) and *very likely* (MIROC5 and MIROC-ES2L) confidence intervals, and some have strong high cloud altitude feedbacks that lie above the upper bound of the *likely* (HadGEM2-ES and CanESM5) and *very likely* (E3SM-1-0) confidence intervals. This feedback component has the greatest number of models (3) lying outside of the assessed *very likely* range; these are the same three models that lie outside the assessed *very likely* range for total cloud feedback. Such inter-model variation is striking for a feedback having a strong theoretical basis and both observational and high-resolution modeling support.

Consistent with Klein et al. (2017), the distribution of modeled tropical marine low cloud feedback values favors the low end of the expert assessed value. No models exceed the central expert assessed value, and several models' values lie below the lower bound of the *likely* (MIROC5, MRI-CGCM3, MIROC-ES2L, and MIROC6) and *very likely* (CCSM4) confidence intervals.

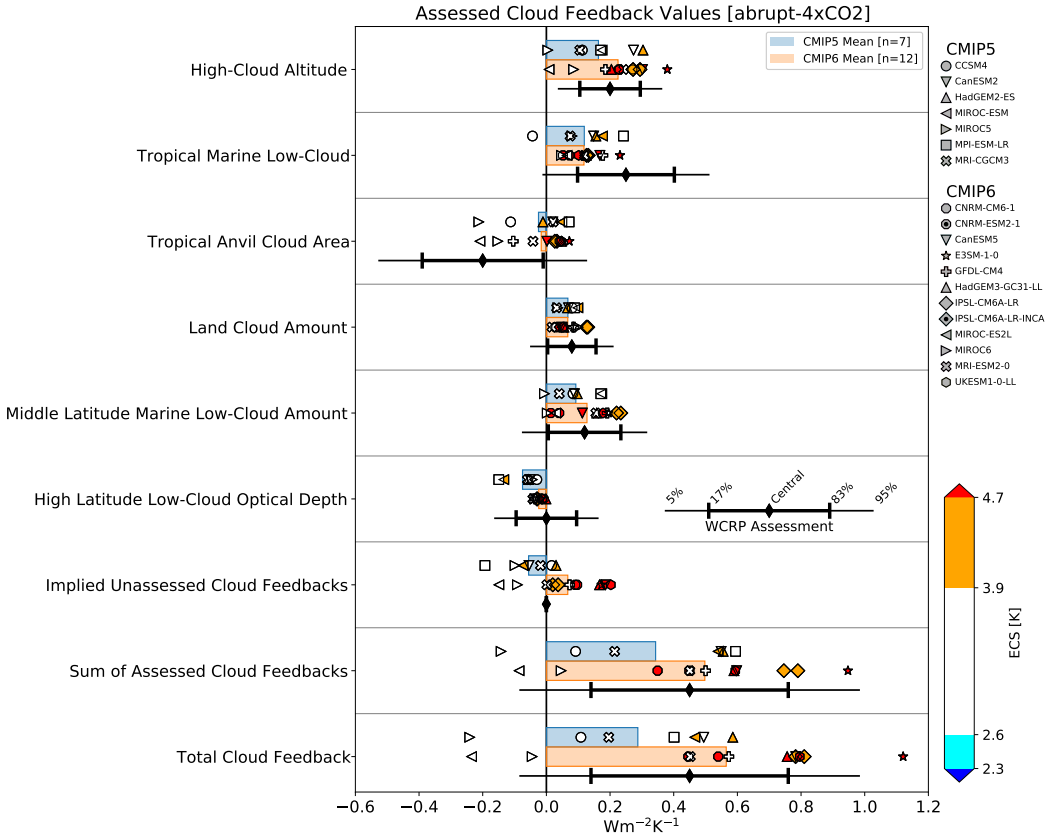

Figure 1. Cloud feedback components estimated from climate model simulations and as assessed in Sherwood et al. (2020). For each component, the individual model values are indicated with symbols, the multi-model means are indicated with blue (CMIP5) and orange (CMIP6) bars, and the expert assessed *likely* and *very likely* confidence intervals are indicated with black errorbars. Model symbols are color-coded by ECS with color boundaries corresponding to the edges of the *likely* and *very likely* ranges of the Baseline posterior PDF of ECS from Sherwood et al. (2020). Identical figures highlighting each individual model are provided in Figures S1-S19.

In contrast, all but two models (MIROC5 and MIROC-ES2L) underestimate the strength of the negative anvil cloud feedback as assessed in Sherwood et al. (2020). Nine models have positive anvil feedbacks that place them above the upper bound of the assessed *likely* confidence inerval (CanESM2, MIROC-ESM, MPI-ESM-LR, MRI-CGCM3, CanESM5, E3SM-1-0, HadGEM3-GC31-LL, IPSL-CM6A-LR, and UKESM1-0-LL).

All models lie within the assessed *likely* range for the land cloud amount feedback and all but two models (MIROC5 and MIROC6) lie within the assessed *likely* range of the middle latitude marine low cloud amount feedback. These two models have values just below the lower bound of the assessed *likely* confidence inerval.

Whereas the central estimate of the high latitude low cloud optical depth feedback from the assessment is 0, all models simulate a negative feedback. All but two models (MIROC-ESM and MPI-ESM-LR) fall within the *likely* assessed range, however. In the multi-model average, the negative feedback values are more than halved in CMIP6 relative to CMIP5, bringing CMIP6 models into better agreement with expert judgement. This may be related to a weakened cloud phase feedback owing to improved simulation of mean-state cloud phase (Bodas-Salcedo et al., 2019; Gettelman et al., 2019; Zelinka et al., 2020; Flynn & Mauritsen, 2020). The inter-model spread in this feedback component has also dramatically decreased.

The unassessed feedback is near zero on average across all models, consistent with it being assigned a value of zero in the expert assessment. However, its across-model standard deviation and its CMIP5-to-CMIP6 increase in multi-model average are greater than for any of the previously discussed individual cloud feedback components. Contributors to this feedback will be discussed in greater detail in Section 3.5.

The sum of all six assessed feedback components is positive in all but two models (MIROC5 and MIROC-ES2L) and exhibits substantially more inter-model spread than any individual component comprising it. Its standard deviation ($\sigma = 0.30$ Wm^{-2}K^{-1}) is also larger than would exist if the feedback components comprising it were uncorrelated across models (σ if summing individual uncertainties in quadrature = 0.21 Wm^{-2}K^{-1}), as discussed further in Section 3.2. While the multi-model mean value is close to the expert-assessed value, some models lie below the lower bound of the assessed *likely* (CCSM4 and MIROC6) and *very likely* (MIROC5 and MIROC-ES2L) confidence intervals, and two lie above the upper bound of the assessed *likely* confidence interval (E3SM-1-0 and IPSL-CM6A-LR).

The total cloud feedback, which is the sum of assessed and unassessed components, has a larger standard deviation than would occur if these two components were uncorrelated. Owing to this correlation, all but four models (CCSM4, CanESM2, MIROC-ESM, and MPI-ESM-LR) exhibit degraded agreement with expert assessment once accounting for their unassessed feedbacks. In addition to the models that fell outside the *likely* and *very likely* ranges for the sum of assessed feedbacks, there are two additional models (CanESM5 and UKESM1-0-LL) that now lie above the upper bound of the assessed *likely* confidence interval, and E3SM-1-0 has now moved above the upper bound of the assessed *very likely* confidence interval.

Unsurprisingly, models with larger total cloud feedback tend to have higher ECS. All models with total cloud feedbacks above the upper limit of the expert-assessed *likely* range are part of CMIP6. These models also have ECS values above 3.9 K, the upper limit of the expert-assessed *likely* ECS range, and three of them have ECS values above 4.7 K, the upper limit of the *very likely* ECS range. However, two models with ECS $>$ 3.9 K (HadGEM2-ES, MIROC-ESM) and even three with ECS $>$ 4.7 K (CNRM-CM6-1, CNRM-ESM2-1, and HadGEM3-GC31-LL) have total cloud feedbacks within the *likely* range, indicating that other non-cloud feedbacks are pushing them to very high ECS. No models considered here have ECS values below 2.6 K, the lower limit of the Sherwood

et al. (2020) assessed *likely* range. Even the models whose cloud feedbacks lie below the lower limit of the *likely* and *very likely* total cloud feedback confidence bound still lie within the *likely* ECS range.

Turning now to the multi-model mean cloud feedback components, we see that the mean total cloud feedback is roughly twice as large in CMIP6 than in CMIP5, qualitatively consistent with Zelinka et al. (2020), who assessed a much larger collection of models. This occurs because the high cloud altitude, midlatitude marine low cloud amount, high latitude low cloud optical depth, and unassessed feedbacks all become more positive, on average, in CMIP6. The other feedbacks remain unchanged on average. The largest shift among individual components is for the unassessed feedback, discussed further in Section 3.5.

All multi-model mean assessed feedback components lie within the respective expert-assessed *likely* range. They also lie very close to the central assessed values, with two exceptions: The tropical marine low cloud feedback averaged across all models (0.12 ± 0.07 $\mathrm{Wm^{-2}K^{-1}}$) is about half as large as assessed (0.25 ± 0.16 $\mathrm{Wm^{-2}K^{-1}}$), and the tropical anvil cloud area feedback averaged across all models is close to zero (-0.02 ± 0.09 $\mathrm{Wm^{-2}K^{-1}}$), whereas it was assessed to be moderately negative (-0.20 ± 0.20 $\mathrm{Wm^{-2}K^{-1}}$). For these two components, GCM values were not used to inform the expert judgement value, but rather they were based upon observations and, in the case of tropical marine low cloud feedbacks, large eddy simulations that resolve many of the cloud processes that must be parameterized in GCMs (see Table 1 of Sherwood et al. (2020)).

3.2 Correlations Among GCM Cloud Feedbacks

The previous section provided several indications that models with large positive total cloud feedbacks tend to have systematically higher cloud feedbacks for *all* components rather than having a single anomalously strong positive component, and vice versa for models with small or negative total cloud feedbacks. We quantify this more rigorously in this section by diagnosing the correlation structure among the individual components.

All individual cloud feedback components are positively correlated with the total cloud feedback, especially the high cloud altitude, tropical anvil, midlatitude marine low cloud amount, and unassesed feedbacks (Figure 2a, column 1). The high cloud altitude feedback has a larger correlation than any other single component. While the tropical marine low cloud feedback is significantly correlated with the total, it is markedly weaker than for several other components, which is surprising given previous findings that low latitude marine low clouds in regions of moderate subsidence drive inter-model spread in climate sensitivity (Bony & Dufresne, 2005). The discrepancy may arise from the relatively small subset of models considered here.

The positive correlations between individual components and the total cloud feedback is expected: If all the models were distributed randomly for each feedback component, one would expect the models with largest total cloud feedback to be the ones that most consistently lie on the positive tail of all components. To demonstrate this, we generated normal distributions with 10,000 samples matching the multi-model mean and standard deviation for each of the six assessed and one unassessed components and repeated the above calculations on these random data. All individual components are significantly positively correlated with their sum, with correlation strengths proportional to the individual component variances (Figure 2b, column 1).

The prevalence of strong and significant positive correlations among individual feedback components seen in the actual model data is, however, not expected from chance. This leads to (1) individual components being more strongly correlated with the total cloud feedback and (2) a wider spread in the total cloud feedback than would occur if

Figure 2. Matrix showing the across-model correlation among all cloud feedback components for (a) actual model data and (b) synthetic normally-distributed data with means and standard deviations equal to those of the models for each feedback component. Correlations that are significantly different from zero at the 95% confidence level are indicated with an asterisk.

individual components were uncorrelated. Models with large positive total cloud feedbacks tend to have systematically larger-than-average cloud feedbacks across multiple components rather than being generally near-average but having a single large component. E3SM-1-0, for example, has the largest positive total cloud feedback, and its feedback values are among the three largest in all categories except the land cloud feedback (Figure S11 and Table 2). Conversely, models like MIROC5 with negative total cloud feedbacks tend to have cloud feedbacks on the left tail of the distribution for *all* components (Figure S5 and Table 2). Consistent with this, we find that most models with near-average total cloud feedbacks have components that are systematically near-average rather than having several components with extreme values of opposing sign that counter each other. One exception is MPI-ESM-LR, which has feedbacks on the high tail of the model distribution for some components and on the low tail for others (Figure S6 and Table 2).

One noteworthy, albeit insignificant, negative correlation in Figure 2a is between the tropical marine low cloud feedback and high latitude low cloud optical depth feedback ($r = -0.45$). This anti-correlation is qualitatively consistent with that shown in Figure 3b of Zelinka et al. (2016) and may be traced to compensating errors in how cloud micro- and macrophysical properties are simulated (McCoy et al., 2016).

That all of the *significant* correlations in Figure 2a are positive might suggest that they are linked by a physical mechanism rather than arising from tuning artifacts. Moreover, several of the significant correlations involve the same cloud types. For example, one could consider the tropical and middle latitude marine low-cloud feedbacks and the (low-cloud dominated) land cloud amount feedback in one grouping, and the high-cloud altitude, anvil, and unassessed feedbacks in another grouping. (As shown in Section 3.5, the unassesed feedbacks are dominated by extra-tropical high-cloud amount and optical depth components.) Given that they involve similar cloud types, it is plausible that positive correlations within these groupings reflect a shared physical mechanism. Other large positive correlations (e.g., between high-cloud altitude and tropical and middle latitude marine low cloud amount) are harder to rationalize. We discuss further implications of all of these correlations in Section 4.

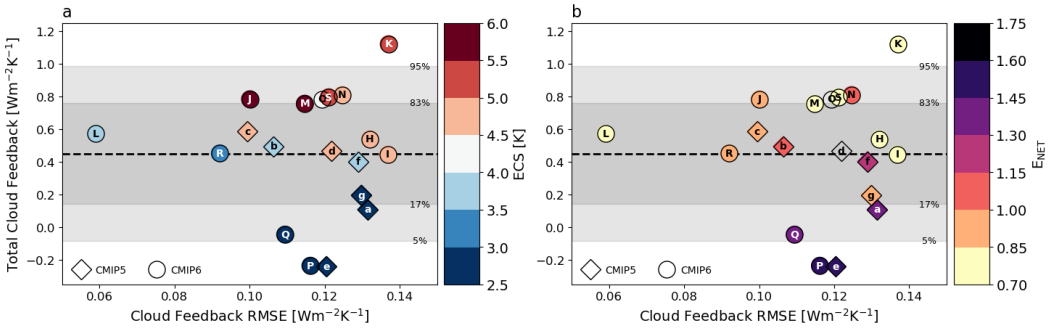

Figure 3. Total cloud feedback scattered against cloud feedback RMSE, with expert *likely* and *very likely* ranges of total cloud feedback indicated with horizontal shading. Models are denoted by letters listed in Table 2 and are colored according to their (a) ECS values and (b) net radiatively-relevant cloud property error metric.

3.3 Metrics of Overall Cloud Feedback Errors

To assess the overall skill of each model in matching the expert-assessed cloud feedback components, we compute a single cloud feedback error metric for each model as the root mean square error (RMSE) with respect to the central expert judgement value over all six assessed feedback components of Sherwood et al. (2020). Each model's cloud feedback RMSE is provided in Table 2 and is plotted against total cloud feedback in Figure 3.

CMIP5 and CMIP6 models exhibit both high and low cloud feedback RMSE values, and the multi-model mean RMSE is only slightly smaller in CMIP6 than in CMIP5 (Table 2). Although the two best-performing models in this measure are CMIP6 models, there is no systematic tendency for CMIP6 models to be performing better than CMIP5 models with respect to expert judgement. For models from the same modelling centers that can be tracked between the two generations, more models show improved performance than degraded performance in this measure: CanESM5 [J] has lower RMSE than its predecessor (CanESM2 [b]); MRI-ESM2-0 [R] has lower RMSE than its predecessor (MRI-CGCM3 [g]); MIROC-ES2L [P] has lower RMSE than its predecessor (MIROC-ESM [d]); and MIROC6 [Q] has lower RMSE than its predecessor (MIROC5 [e]); while the two UKMO models (HadGEM3-GC31-LL [M] and UKESM1-0-LL [S]) have higher RMSE than their predecessor (HadGEM2-ES [c]).

The four models with smallest cloud feedback errors (i.e., RMSE $\leq$ 0.10 Wm^{-2}K^{-1}) have moderate (0.4–0.6 Wm^{-2}K^{-1}) total cloud feedbacks and moderate (3–4 K) ECS values, except for HadGEM2-ES [c] which has an ECS of 4.6 K. This makes sense given that the expert-assessed value of total cloud feedback, which has the greatest leverage on ECS, led to moderate values of ECS in Sherwood et al. (2020). GFDL-CM4 [L] has the lowest RMSE of all models, followed by MRI-ESM2-0 [R]. These are the only models for which all assessed feedbacks lie within the expert *likely* range, and most of their feedback components lie near the assessed central values (Figure S12 and S18; Table 2). Put simply, they get the right answer for the right reasons.

Models with too-large or too-small total cloud feedbacks and ECS tend to have larger-than-average cloud feedback RMSE values. That is, the models that lie farthest from the horizontal dashed line tend to be located on the right side of Figure 3. The models with small total cloud feedback and small ECS have cloud feedback components that are systematically biased low relative to expert judgement, giving them larger-than-average RMSE. Of the five models with ECS $<$ 3 K, only MIROC6 [Q] has a below-average RMSE

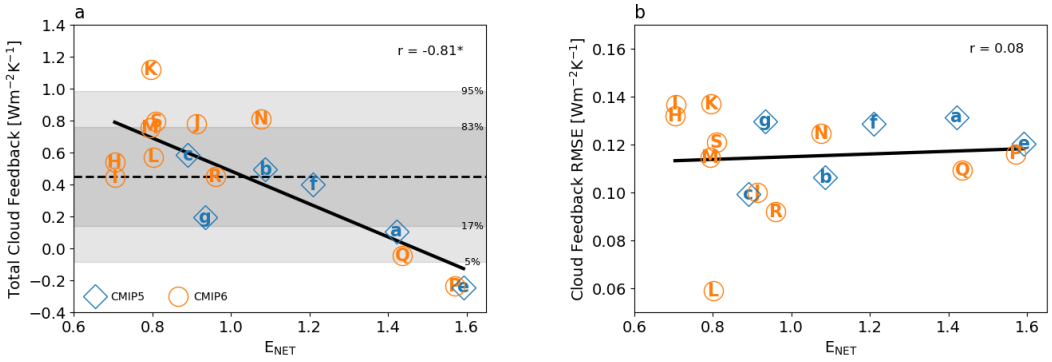

Figure 4. (a) Total cloud feedback and (b) cloud feedback RMSE scattered against net radiatively-relevant cloud property error metric. Models are denoted by letters listed in Table 2 and are colored blue for CMIP5 and orange for CMIP6. Expert *likely* and *very likely* ranges of total cloud feedback indicated with horizontal shading in (a). Correlations that are significant at 95% confidence are indicated with an asterisk.

value. The models with large total cloud feedback and large ECS have cloud feedback components that are systematically biased high relative to expert judgement, also giving them larger-than-average RMSE. Of the nine models with ECS > 4.5 K, only HadGEM2-ES [c], CanESM5 [J], and HadGEM3-GC31-LL [M] have below-average RMSE value. E3SM-1-0 [K] has the highest RMSE of all models considered. Although it lies within the assessed *likely* range for four components, it lies above the upper limit of the assessed *likely* range for two components (Figure S11; Table 2).

However, several models with total cloud feedbacks very close to the central value of expert assessment have moderate and even large RMSE values. Most notably, CNRM-ESM2-1 [I] achieves a reasonable total cloud feedback partly through having a low-biased tropical marine low cloud feedback that counteracts its high-biased tropical anvil cloud area feedback (Figure S9; Table 2). Put simply, it gets the right answer for the wrong reasons.

GFDL-CM4, MRI-ESM2-0, HadGEM2-ES, and CanESM2 remain among the five models with lowest RMSE regardless of whether we use feedbacks derived from `abrupt-4xCO2` or `amip-p4K` experiments.

3.4 Relationship Between Cloud Feedbacks and Mean-State Cloud Property Errors

The fidelity with which models simulate mean-state radiatively-relevant cloud properties is strongly and significantly correlated with total cloud feedback (Figure 4a). We show this result for the net radiatively-relevant cloud property error, but it is also strong and significant for the SW-radiation error as well as the cloud property error without radiative weighting (not shown). This result is consistent with Figure 11 of Klein et al. (2013), but now the relationship holds across two ensembles of models (CMIP5 and CMIP6). It is possible that the CFMIP1 models were offset in Klein et al. (2013) because mean-state cloud errors are larger in non-AMIP simulations where SSTs can deviate substantially from those observed in nature. However, we find little support for this, as error metrics derived in `amip` and `piControl` experiments are virtually indistinguishable (not shown). While caution is necessary given the relatively small sample size, an important question is why better simulating present-day cloud properties is associated with larger cloud feedbacks. We leave this as an open question for future research.

On average, mean-state cloud properties are simulated better in CMIP6 than in CMIP5 (Figure 4a; Table 2). Six CMIP6 models now have smaller error values than the smallest exhibited in CMIP5. For models from the same modeling center than can be tracked, all but one has improved in this measure from CMIP5 to CMIP6. Specifically, marked improvement is seen from CanESM2 [b] to CanESM5 [J], from HadGEM2-ES [c] to HadGEM3-GC31-LL [M] and UKESM1-0-LL [S], and from MIROC5 [e] to MIROC6 [Q], whereas MRI-ESM2-0 [R] has very slightly degraded mean-state clouds relative to MRI-CGCM3 [g].

It is often implicitly assumed by model developers and model analysts that the degree to which a model's clouds resembles reality can be used as a basis to trust their response to climate change. In Figure 4b, we test this assumption by comparing the agreement with expert judgment for cloud feedbacks to the agreement with observations of the present-day climatological distribution of clouds and their properties. The absence of any relationship between these two metrics clearly indicates that improved simulation of mean-state cloud properties does not lead to improved cloud feedbacks with respect to expert judgment. This can also be seen in Figure 3b, where models are color-coded by E_{NET}, allowing for a simultaneous assessment of how well models simulate mean-state cloud properties and match expert judgment of total cloud feedback and its components. From this it is evident that most of the models with small mean-state errors (yellow shading) have large cloud feedback RMSE and several lie above the upper limit of the *likely* range of total cloud feedback (i.e., in the top-right portion of the diagram). The one exception is GFDL-CM4 [L], which achieves low cloud feedback RMSE, low values of E_{NET}, and total cloud feedback near the central value of expert judgement.

At the same time, simulating worse-than-average mean-state cloud properties is generally associated with poorer agreement with the expert-assessed total cloud feedback and/or its components. This is evidenced by the fact that most of the models with large mean-state errors (purple/black shading) have moderate cloud feedback RMSE and lie below the lower limit of the *likely* range of total cloud feedback (i.e., in the bottom-right part of Figure 3b). MPI-ESM-LR [f], which achieves a reasonable total cloud feedback despite its large feedback RMSE, has a larger E_{NET} than other models that lie near the expert assessed total cloud feedback value.

3.5 GCM Cloud Feedbacks in Unassessed Categories

Figure 5 shows a breakdown of explicitly-computed feedbacks that were not assessed in Sherwood et al. (2020). Examining these feedback components is important as it may guide where future research with observations, process-resolving models, and theory is needed to further constrain GCMs' cloud feedbacks. There are an infinite number of ways of breaking down these components, but our strategy was to quantify those that complement the assessed feedbacks, either in altitude or geographic space, to the extent possible. For example, we quantify the low cloud altitude feedback since the high cloud feedback is an assessed category, and we quantify the low cloud optical depth feedback between 30 and 90 degrees latitude but excluding the 40–70 degree zone where it was already assessed. The sum of these closely reproduces the implied unassessed feedbacks in Figure 1 (not shown).

The multi-model mean unassessed cloud feedback transitions from being -0.06 $Wm^{-2}K^{-1}$ on average in CMIP5 to $+0.06$ $Wm^{-2}K^{-1}$ on average in CMIP6. It becomes more positive for all individual components except the (very small) low cloud altitude feedback. The largest shifts occur for the multi-model mean extratropical high cloud optical depth and amount components, which both transition from negative to positive values.

There are a few models whose unassessed feedbacks sum to a value that is large relative to their total and/or combined assessed feedbacks and worth examining in greater detail. Strong negative unassessed cloud feedbacks (with values < -0.10 $Wm^{-2}K^{-1}$) are

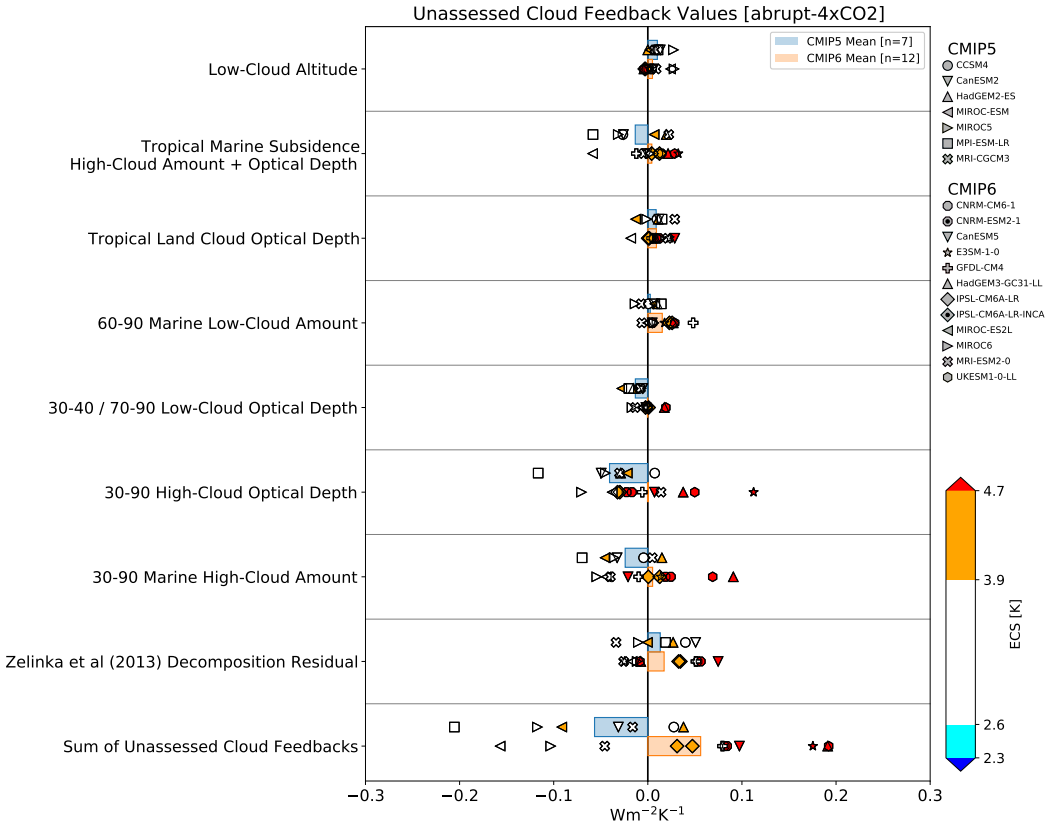

Figure 5. As in Figure 1, but for cloud feedback components that were not assessed in Sherwood et al. (2020). Note the x-axis spans a range that is only a third of that in Figure 1.

around half as large in magnitude as the total cloud feedbacks in MPI-ESM-LR and MIROC5, and are actually larger in magnitude than the sum of assessed feedbacks in MIROC-ES2L and MIROC6. All of these models have negative high cloud feedbacks in common. Most notably, all four have negative extratropical high cloud optical depth and amount components with values < -0.04 $\mathrm{Wm^{-2}K^{-1}}$. Additionally, all but MIROC6 have moderately negative feedbacks for high cloud amount and optical depth in tropical marine subsidence regions.

Anomalously large positive extratropical high cloud feedbacks also cause large positive unassessed feedbacks (exceeding $+0.17$ $\mathrm{Wm^{-2}K^{-1}}$) in three models: E3SM-1-0's extratropical high cloud optical depth feedback is much larger than in any other model, while in HadGEM3-GC31-LL and UKESM1-0-LL, this feedback along with an even stronger extratropical high cloud amount feedback are jointly responsible.

4 Discussion and Conclusions

We have evaluated cloud feedback components simulated in 19 CMIP5 and CMIP6 models against benchmark values determined via an expert synthesis of observational, theoretical, and high-resolution modeling studies (Sherwood et al., 2020). We found that models that most closely match the expert-assessed values across several cloud feedback components have moderate total cloud feedbacks (0.4–0.6 $\mathrm{Wm^{-2}K^{-1}}$) and generally moderate ECS (3–4 K). In contrast, models with largest feedback errors with respect to expert assessment generally have total cloud feedbacks and climate sensitivities that are too large or too small.

There is no evidence that CMIP6 models simulate cloud feedbacks in better agreement with expert judgement than do CMIP5 models. While the two best models in our error metric are CMIP6 models, more modeling centers show degraded than improved performance between their CMIP5 and CMIP6 model versions. All models with total cloud feedbacks above the upper limit of the expert-assessed *likely* range are part of CMIP6 and have ECS values above 3.9 K, the upper limit of the expert-assessed *likely* ECS range. However, the converse is not true: several models with high ECS have total cloud feedbacks within the *likely* range. This means that large cloud feedback ensures a high ECS, but high ECS can emerge even with moderate cloud feedbacks, a result consistent with M. J. Webb et al. (2013) for CMIP3 models. More generally, having $2\mathrm{xCO_2}$ radiative forcing and feedbacks in agreement with expert judgement does not guarantee that a model's ECS will be in agreement with expert judgement because the latter is further constrained by evidence from the paleoclimate and historical records (Sherwood et al., 2020).

On average, and for most individual modeling centers, mean-state cloud properties are better simulated in CMIP6. Better simulation of mean-state cloud properties is strongly and significantly correlated with larger total cloud feedback, for reasons that remain to be investigated. But more skillful simulation of mean-state cloud properties does not, in general, translate to more skillful simulation of cloud feedbacks, and many models with small mean-state errors have large cloud feedback errors with respect to expert judgment. In general, better simulation of mean-state cloud properties leads to stronger but not necessarily better cloud feedbacks. GFDL-CM4, which has the smallest cloud feedback error, small mean-state cloud property error, and a total cloud feedback near the expert-assessed central value, is the exception to this rule.

Models with large positive total cloud feedbacks tend to have systematically higher cloud feedbacks for all components rather than having a single anomalously strong positive component, and vice versa for models with small or negative total cloud feedbacks. This means, for example, that there is no single feedback that all high ECS models are exaggerating. However, if there is some physical relationship causing the correlation between individual feedback components, this may imply that constraining one component

would have knock-on effects across several components. In this case, feedbacks from multiple cloud types could be constrained with less evidence than would be needed if they were uncorrelated, and changing one aspect of a model might systematically change the feedbacks from multiple cloud types, making it easier to improve its cloud feedbacks.

One plausible scenario could involve systematically more positive feedback components in models with optically thicker mean-state clouds. Models with thicker clouds would be expected to have larger positive amount components for a given decrease in cloud fraction, larger positive altitude feedbacks for a given increase in cloud top altitude, and weaker negative optical depth feedbacks for a given increase in cloud water content. Such a linkage between mean-state cloud properties and multiple feedback components is qualitatively consistent with the strong correlation between skill in simulating mean-state clouds and larger cloud feedback noted above. Establishing and understanding the physical basis of correlations among feedback components and their potential linkages with mean-state cloud properties is important future work.

Taken as a whole, cloud feedbacks in CMIP6 exhibit two noteworthy changes relative to CMIP5. First, the high latitude low-cloud optical depth feedback has shifted from being robustly negative across CMIP5 models, with some models simulating moderately strong negative feedbacks below the expert-assessed *likely* range, to a much weaker negative feedback in CMIP6, with the models tightly clustered about it. This represents a shift towards better agreement with expert judgement, and may be tied to reductions in super-cooled liquid biases in the latest models (Bodas-Salcedo et al., 2019; Gettelman et al., 2019; Zelinka et al., 2020). Second, the inter-model spread in the tropical marine low-cloud feedback has decreased markedly between CMIP5 and CMIP6, with the across-model standard deviation nearly halving. This may indicate some degree of model convergence in the simulated response of tropical low clouds to warming, albeit one that is centered on the lower end of the expert-assessed *likely* range.

Results from several individual cloud feedback components raise important questions and motivate future investigation:

- The high cloud altitude feedback strength varies widely across models, despite its firm theoretical basis and support from observational analyses and high-resolution modeling. This motivates further work to pin down causes of inter-model spread and to eliminate sources of bias in this feedback.
- Although we found that the tropical marine low cloud feedback simulated by most models lies at the low end of the expert-assessed *likely* range, recent observational constraints support slightly lower values (Cesana & Del Genio, 2021) (Myers et al 2021) owing in part to a better discrimination between strong stratocumulus feedbacks and weaker trade cumulus feedbacks. If incorporated into a future assessment, the expert value of this feedback could be revised downward, *likely* resulting in a better alignment between it and the multi-model mean. To the extent that the assessed confidence bounds also narrow, however, the models with very weak tropical marine low cloud feedbacks may still lie below the expert judgement range.
- Despite the wide uncertainty in its expert-assessed value, nine models have positive tropical anvil cloud feedbacks that place them above the upper bound of the assessed *likely* confidence interval. This discrepancy between models and expert judgment can be traced to the disagreement between models and observations in the sensitivity of tropical TOA radiation and deep convective cloud properties to interannual fluctuations in surface temperature found in the studies of (Mauritsen & Stevens, 2015; Williams & Pierrehumbert, 2017), which were influential in establishing the expert-assessed value. Much uncertainty remains surrounding the processes controlling tropical anvil cloud fraction and its changes with warming,

and the fidelity with which GCMs can simulate them (Bony et al., 2016; Hartmann, 2016; Seeley et al., 2019; Wing et al., 2020; Gasparini et al., 2021).

- Cloud feedback components that were not assessed in Sherwood et al. (2020), though summing to zero on average across models, have substantial inter-model spread and play the largest role in the increase in multi-model average cloud feedback from CMIP5 to CMIP6. Of these, extratropical high cloud feedbacks emerge as strong in many models. It remains to be established whether and to what extent these feedbacks are exaggerated in models or rather important components that need to be explicitly included in future assessments. This, along with the aforementioned uncertainties surrounding high cloud altitude and anvil cloud feedbacks highlights the need for further observational analyses, process-resolving modeling, and theoretical studies targeting high cloud feedbacks.

We have provided Python code that performs all calculations and generates all visualizations presented in this study. The code is also easily modified to accommodate comparisons between GCM cloud feedbacks and the similar but not identical breakdown of cloud feedback components that is used in the forthcoming 6th Assessment report of the IPCC. We envision that this code could be applied to perturbed parameter or perturbed physics ensembles and to developmental versions of models to assess cloud feedbacks and cloud errors and place them in the context of other models and of expert judgement in real-time during model development. This may be particularly valuable in less computationally expensive prescribed SST perturbation experiments that are routinely performed during model development. Despite their simpler design, these "Cess-type" experiments effectively capture the feedbacks present in fully coupled experiments (Ringer et al., 2014).

Acknowledgments

Python code to perform all calculations and produce all figures and tables in this manuscript is available at `https://github.com/mzelinka/assessed-cloud-fbks`. CMIP5 and CMIP6 ECS values are available at `https://github.com/mzelinka/cmip56_forcing_feedback_ecs`. ISCCP HGG cloud data is provided by NOAA/NCEI at `https://www.ncei.noaa.gov/thredds/catalog/cdr/isccp_hgg_agg/files/catalog.html`. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP and ESGF. This work was supported by the U.S. Department of Energy (DOE) Regional and Global Modeling Analysis program area and was performed under the auspices of the DOE by Lawrence Livermore National Laboratory under Contract DEAC5207NA27344. We are grateful for stimulating discussions with Leo Donner, Chris Golaz, Tim Myers, Yoko Tsushima, and Mark Webb.

References

Bodas-Salcedo, A., Mulcahy, J. P., Andrews, T., Williams, K. D., Ringer, M. A., Field, P. R., & Elsaesser, G. S. (2019). Strong Dependence of Atmospheric Feedbacks on Mixed-Phase Microphysics and Aerosol-Cloud Interactions in HadGEM3. *Journal of Advances in Modeling Earth Systems*, *11*(6), 1735–1758. doi: 10.1029/2019ms001688

Bodas-Salcedo, A., Webb, M. J., Bony, S., Chepfer, H., Dufresne, J. L., Klein, S. A., ... John, V. O. (2011). COSP Satellite simulation software for model assessment. *Bulletin of the American Meteorological Society*, *92*(8), 1023–1043. doi: 10.1175/2011bams2856.1

Bony, S., & Dufresne, J. L. (2005). Marine boundary layer clouds at the heart of

Table 1. Central value and 1-σ uncertainty of the cloud feedback components assessed in Sherwood et al. (2020) (in $\mathrm{Wm^{-2}K^{-1}}$), and description of how each component is computed in GCMs in this study.

Expert-Assessed Feedbacks		Calculation in GCMs			
Name	Value	Components	Surface	Regime	Region
1. high cloud altitude	0.2 ± 0.10	high cloud altitude	all	all	90S-90N
2. tropical marine low-cloud	0.25 ± 0.16	sum of low cloud amount, altitude, & optical depth	ocean	descent	30S-30N
3. tropical anvil cloud area	-0.2 ± 0.20	sum of all components except high cloud altitude	ocean	ascent	30S-30N
4. land cloud amount	0.08 ± 0.08	sum of low and high cloud amount	land	all	90S-90N
5. middlelatitude marine lowcloud amount	0.12 ± 0.12	low cloud amount	ocean	all	30-60N/S
6. highlatitude lowcloud optical depth	0.00 ± 0.10	low cloud optical depth	all	all	40-70N/S
7. sum of assessed	0.45 ± 0.33	sum of items 1–6			
8. total cloud feedback	0.45 ± 0.33	total cloud feedback	all	all	90S-90N
9. implied unassessed	N/A	item 8 minus item 7			

Table 2. Individual cloud feedback components (in $Wm^{-2}K^{-1}$), cloud feedback RMSE values (in $Wm^{-2}K^{-1}$), net radiatively-relevant cloud property error metrics (E_{NET}; unitless), and effective climate sensitivities (ECS; K) for all models analyzed in this study. Expert-assessed central values and uncertainties of cloud feedback components are also provided. Any model values that lie outside of the *very likely* (90%) and *likely* (66%) confidence intervals of expert judgement are denoted with double and single asterisks, respectively.

Model	Variant	High Alt.	Marine Low	Tropical Anvil	Land Amt.	Midlat Low Amt.	Hilat Low Tau	Unassessed	Sum Assessed	Total	RMSE	E_{NET}	ECS
a) CCSM4	r2i1p1	0.11	-0.04**	-0.11	0.08	0.08	-0.03	0.02	0.09*	0.11*	0.13	1.42	2.94
b) CanESM2	r1i1p1	0.27	0.15	0.02*	0.07	0.09	-0.05	-0.05	0.55	0.49	0.11	1.09	3.70
c) HadGEM2-ES	r1i1p1	0.30*	0.16	-0.01	0.07	0.10	-0.06	0.03	0.56	0.59	0.10	0.89	4.60*
d) MIROC-ESM	r1i1p1	0.17	0.18	0.05*	0.10	0.17	-0.13*	-0.07	0.54	0.47	0.12	N/A	4.65*
e) MIROC5	r1i1p1	0.00**	0.08*	-0.21	0.04	-0.01*	-0.04	-0.10	-0.14**	-0.24**	0.12	1.59	2.71
f) MPI-ESM-LR	r1i1p1	0.17	0.24	0.07*	0.09	0.17	-0.15*	-0.19	0.59	0.40	0.13	1.21	3.63
g) MRI-CGCM3	r1i1p1	0.10*	0.07*	0.02*	0.03	0.04	-0.06	-0.02	0.21	0.20	0.13	0.93	2.61
CMIP5 Average		0.16	0.12	-0.02	0.07	0.09	-0.07	-0.06	0.34	0.29	0.12	1.19	3.55
CMIP5 1-σ		0.10	0.09	0.09	0.02	0.06	0.04	0.07	0.27	0.27	0.01	0.25	0.78
H) CNRM-CM6-1	r1i1p1f2	0.27	0.07*	0.04*	0.03	0.04	-0.01	0.09	0.45	0.54	0.13	0.70	4.90**
I) CNRM-ESM2-1	r1i1p1f2	0.23	0.05*	0.04*	0.03	0.01	-0.01	0.10	0.35	0.45	0.14	0.71	4.79**
J) CanESM5	r1i1p2f1	0.30*	0.17	0.00*	0.05	0.11	-0.03	0.19	0.60	0.78*	0.10	0.91	5.62**
K) E3SM-1-0	r1i1p1f1	0.38**	0.23	0.07*	0.09	0.19	-0.02	0.17	0.95*	1.12**	0.14	0.80	5.31**
L) GFDL-CM4	r1i1p1f1	0.19	0.18	-0.10	0.09	0.19	-0.03	0.07	0.50	0.57	0.06	0.80	3.89
M) HadGEM3-GC31-LL	r1i1p1f3	0.20	0.13	0.05*	0.06	0.16	-0.00	0.17	0.59	0.76	0.11	0.79	5.55**
N) IPSL-CM6A-LR	r1i1p1f1	0.29	0.13	0.03*	0.13	0.23	-0.03	0.02	0.79*	0.81*	0.12	1.08	4.70*
O) IPSL-CM6A-LR-INCA	r1i1p1f1	0.27	0.13	0.03*	0.13	0.22	-0.03	0.04	0.75	0.78*	0.12	N/A	4.13*
P) MIROC-ES2L	r1i1p1f2	0.01**	0.06*	-0.21	0.05	0.03	-0.03	-0.15	-0.09**	-0.23**	0.12	1.57	2.66
Q) MIROC6	r1i1p1f1	0.09*	0.05*	-0.15	0.10	0.00*	-0.04	-0.09	0.05*	-0.04*	0.11	1.44	2.60
R) MRI-ESM2-0	r1i1p1f1	0.24	0.12	-0.04	0.02	0.16	-0.04	0.00	0.45	0.45	0.09	0.96	3.13
S) UKESM1-0-LL	r1i1p1f2	0.23	0.10	0.05*	0.05	0.18	-0.01	0.20	0.59	0.80*	0.12	0.81	5.36**
CMIP6 Average		0.23	0.12	-0.02	0.07	0.13	-0.02	0.07	0.50	0.57	0.11	0.96	4.39
CMIP6 1-σ		0.09	0.05	0.09	0.04	0.08	0.01	0.11	0.28	0.36	0.02	0.28	1.05
CMIP5/6 Average		0.20	0.12	-0.02	0.07	0.11	-0.04	0.02	0.44	0.46	0.12	1.04	4.08
CMIP5/6 1-σ		0.10	0.07	0.09	0.03	0.08	0.04	0.11	0.29	0.36	0.02	0.29	1.04
WCRP Central		0.2	0.25	-0.2	0.08	0.12	0.0	N/A	0.45	0.45			
WCRP 1-σ		0.10	0.16	0.20	0.08	0.12	0.10	N/A	0.33	0.33			

tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.*, *32*. doi: 10.1029/2005GL023851

Bony, S., Dufresne, J. L., Treut, H. L., Morcrette, J. J., & Senior, C. (2004). On dynamic and thermodynamic components of cloud changes. *Climate Dyn.*, *22*, 71–68. doi: 10.1007/s00382-003-0369-6

Bony, S., Stevens, B., Coppin, D., Becker, T., Reed, K. A., Voigt, A., & Medeiros, B. (2016). Thermodynamic control of anvil cloud amount. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1601472113

Cesana, G. V., & Del Genio, A. D. (2021, March). Observational constraint on cloud feedbacks suggests moderate climate sensitivity. *Nature Climate Change*, *11*(3), 213–218. Retrieved 2021-03-09, from `https://www.nature.com/articles/s41558-020-00970-y` doi: 10.1038/s41558-020-00970-y

Cess, R. D., & others. (1990). Intercomparison and interpretation of cloud-climate feedback processes in nineteen atmospheric general circulation models. *J. Geophys. Res.*, *95*, 16601–16615.

Cess, R. D., Potter, G. L., Blanchet, J. P., Boer, G. J., Ghan, S. J., Kiehl, J. T., ... Yagai, I. (1989). Interpretation of Cloud-Climate Feedback as Produced by 14 Atmospheric General Circulation Models. *Science*, *245*(4917), 513–516. doi: 10.1126/science.245.4917.513

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., ... Wehner, M. (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*. Cambridge, United Kingdom and New York, NY, USA.: Cambridge University Press.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, *9*(5), 1937–1958. doi: 10.5194/gmd-9-1937-2016

Flynn, C. M., & Mauritsen, T. (2020, July). On the climate sensitivity and historical warming evolution in recent coupled model ensembles. *Atmospheric Chemistry and Physics*, *20*(13), 7829–7842. Retrieved 2021-03-10, from `https://acp.copernicus.org/articles/20/7829/2020/` doi: https://doi.org/10.5194/acp-20-7829-2020

Gasparini, B., Rasch, P. J., Hartmann, D. L., Wall, C. J., & Dtsch, M. (2021). A Lagrangian Perspective on Tropical Anvil Cloud Lifecycle in Present and Future Climate. *Journal of Geophysical Research: Atmospheres*, *126*(4), e2020JD033487. Retrieved 2021-03-25, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD033487` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020JD033487) doi: https://doi.org/10.1029/2020JD033487

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., ... Mills, M. J. (2019). High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2). *Geophysical Research Letters*, *46*(14), 8329–8337. doi: 10.1029/2019gl083978

Hartmann, D. L. (2016, August). Tropical anvil clouds and climate sensitivity. *Proceedings of the National Academy of Sciences*. Retrieved 2021-03-25, from `https://www.pnas.org/content/early/2016/07/29/1610455113` (Publisher: National Academy of Sciences Section: Commentary) doi: 10.1073/pnas.1610455113

Klein, S. A., Hall, A., Norris, J. R., & Pincus, R. (2017). Low-Cloud Feedbacks from Cloud-Controlling Factors: A Review. *Surveys in Geophysics*. doi: 10.1007/s10712-017-9433-3

Klein, S. A., & Jakob, C. (1999). Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Weath. Rev.*, *127*, 2514–2531. doi: 10

.1175/1520-0493(1999)1272.0.CO;2

Klein, S. A., Zhang, Y., Zelinka, M. D., Pincus, R., Boyle, J., & Gleckler, P. J. (2013). Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator. *Journal of Geophysical Research-Atmospheres*, *118*(3), 1329–1342. doi: 10.1002/jgrd.50141

Mauritsen, T., & Stevens, B. (2015). Missing iris effect as a possible cause of muted hydrological change and high climate sensitivity in models. *Nature Geosci*, *8*(5), 346–351. doi: 10.1038/ngeo2414http://www.nature.com/ngeo/journal/v8/n5/abs/ngeo2414.html#supplementary-information

McCoy, D. T., Tan, I., Hartmann, D. L., Zelinka, M. D., & Storelvmo, T. (2016). On the relationships among cloud cover, mixed-phase partitioning, and planetary albedo in GCMs. *Journal of Advances in Modeling Earth Systems*, *8*(2), 650–668. doi: 10.1002/2015ms000589

Nijsse, F. J. M. M., Cox, P. M., & Williamson, M. S. (2020, August). Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models. *Earth System Dynamics*, *11*(3), 737–750. Retrieved 2021-03-19, from `https://esd.copernicus.org/articles/11/737/2020/` (Publisher: Copernicus GmbH) doi: https://doi.org/10.5194/esd-11-737-2020

Pincus, R., Forster, P. M., & Stevens, B. (2016). The Radiative Forcing Model Intercomparison Project (RFMIP): experimental protocol for CMIP6. *Geoscientific Model Development*, *9*(9), 3447–3460. doi: 10.5194/gmd-9-3447-2016

Ringer, M. A., Andrews, T., & Webb, M. J. (2014). Global-mean radiative feedbacks and forcing in atmosphere-only and coupled atmosphere-ocean climate change experiments. *Geophysical Research Letters*, *41*(11), 4035–4042. doi: 10.1002/2014gl060347

Rossow, W. B., & Schiffer, R. A. (1999). Advances in Understanding Clouds from ISCCP. *Bull. Amer. Meteor. Soc.*, *80*(11), 2261–2287. doi: 10.1175/1520-0477(1999)0802.0.CO;2

Seeley, J. T., Jeevanjee, N., Langhans, W., & Romps, D. M. (2019). Formation of Tropical Anvil Clouds by Slow Evaporation. *Geophysical Research Letters*, *46*(1), 492–501. doi: 10.1029/2018GL080747

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., ... Zelinka, M. D. (2020). An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. Retrieved 2021-02-14, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000678` doi: https://doi.org/10.1029/2019RG000678

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020, March). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12), eaaz9549. Retrieved 2021-03-19, from `https://advances.sciencemag.org/content/6/12/eaaz9549` (Publisher: American Association for the Advancement of Science Section: Research Article) doi: 10.1126/sciadv.aaz9549

Webb, M., Senior, C., Bony, S., & Morcrette, J. J. (2001). Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models. *Climate Dyn.*, *17*, 905–922. doi: 10.1007/s003820100157

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., ... Watanabe, M. (2017). The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384. doi: 10.5194/gmd-10-359-2017

Webb, M. J., Lambert, F. H., & Gregory, J. M. (2013). Origins of differences in climate sensitivity, forcing and feedback in climate models. *Climate Dynamics*, *40*(3-4), 677–707. doi: 10.1007/s00382-012-1336-x

Williams, I. N., & Pierrehumbert, R. T. (2017). Observational evidence against strongly stabilizing tropical cloud feedbacks. *Geophysical Research Letters*, *44*(3), 1503–1510. Retrieved 2021-05-04, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL072202 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016GL072202) doi: https://doi.org/10.1002/2016GL072202

Wing, A. A., Stauffer, C. L., Becker, T., Reed, K. A., Ahn, M.-S., Arnold, N. P., ... Zhao, M. (2020). Clouds and Convective Self-Aggregation in a Multimodel Ensemble of Radiative-Convective Equilibrium Simulations. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002138. Retrieved 2021-03-25, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002138 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002138) doi: https://doi.org/10.1029/2020MS002138

Young, A. H., Knapp, K. R., Inamdar, A., Hankins, W., & Rossow, W. B. (2018, March). The International Satellite Cloud Climatology Project H-Series climate data record product. *Earth System Science Data*, *10*(1), 583–593. Retrieved 2021-03-23, from https://essd.copernicus.org/articles/10/583/2018/ (Publisher: Copernicus GmbH) doi: https://doi.org/10.5194/essd-10-583-2018

Zelinka, M. D., Klein, S. A., & Hartmann, D. L. (2012a). Computing and Partitioning Cloud Feedbacks Using Cloud Property Histograms. Part I: Cloud Radiative Kernels. *Journal of Climate*, *25*(11), 3715–3735. doi: 10.1175/jcli-d-11-00248.1

Zelinka, M. D., Klein, S. A., & Hartmann, D. L. (2012b). Computing and Partitioning Cloud Feedbacks Using Cloud Property Histograms. Part II: Attribution to Changes in Cloud Amount, Altitude, and Optical Depth. *Journal of Climate*, *25*(11), 3736–3754. doi: 10.1175/JCLI-D-11-00249.1

Zelinka, M. D., Klein, S. A., Taylor, K. E., Andrews, T., Webb, M. J., Gregory, J. M., & Forster, P. M. (2013). Contributions of Different Cloud Types to Feedbacks and Rapid Adjustments in CMIP5. *Journal of Climate*, *26*(14), 5007–5027. doi: 10.1175/jcli-d-12-00555.1

Zelinka, M. D., Myers, T. A., McCoy, D. T., PoChedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, *47*(1), e2019GL085782. Retrieved 2020-12-23, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782 doi: https://doi.org/10.1029/2019GL085782

Zelinka, M. D., Zhou, C., & Klein, S. A. (2016). Insights from a refined decomposition of cloud feedbacks. *Geophysical Research Letters*, *43*(17), 9259–9269. doi: 10.1002/2016gl069917

Zhu, J., OttoBliesner, B. L., Brady, E. C., Poulsen, C. J., Tierney, J. E., Lofverstrom, M., & DiNezio, P. (2021). Assessment of Equilibrium Climate Sensitivity of the Community Earth System Model Version 2 Through Simulation of the Last Glacial Maximum. *Geophysical Research Letters*, *48*(3), e2020GL091220. Retrieved 2021-03-19, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091220 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL091220) doi: https://doi.org/10.1029/2020GL091220

Zhu, J., Poulsen, C. J., & Otto-Bliesner, B. L. (2020, May). High climate sensitivity in CMIP6 model not supported by paleoclimate. *Nature Climate Change*, *10*(5), 378–379. Retrieved 2021-01-05, from https://www.nature.com/articles/s41558-020-0764-6 doi: 10.1038/s41558-020-0764-6

Figure 1.

Assessed Cloud Feedback Values [abrupt-4xCO2]

High-Cloud Altitude
Tropical Marine Low-Cloud
Tropical Anvil Cloud Area
Land Cloud Amount
Middle Latitude Marine Low-Cloud Amount
High Latitude Low-Cloud Optical Depth
Implied Unassessed Cloud Feedbacks
Sum of Assessed Cloud Feedbacks
Total Cloud Feedback

CMIP5 Mean [n=7]
CMIP6 Mean [n=12]

CMIP5
- CCSM4
- CanESM2
- HadGEM2-ES
- MIROC-ESM
- MIROC5
- MPI-ESM-LR
- MRI-CGCM3

CMIP6
- CNRM-CM6-1
- CNRM-ESM2-1
- CanESM5
- E3SM-1-0
- GFDL-CM4
- HadGEM3-GC31-LL
- IPSL-CM6A-LR
- IPSL-CM6A-LR-INCA
- MIROC-ES2L
- MIROC6
- MRI-ESM2-0
- UKESM1-0-LL

5% 17% Central 83% 95%
WCRP Assessment

ECS [K]
4.7
3.9
2.6
2.3

−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0 1.2
$Wm^{-2}K^{-1}$

Figure 2.

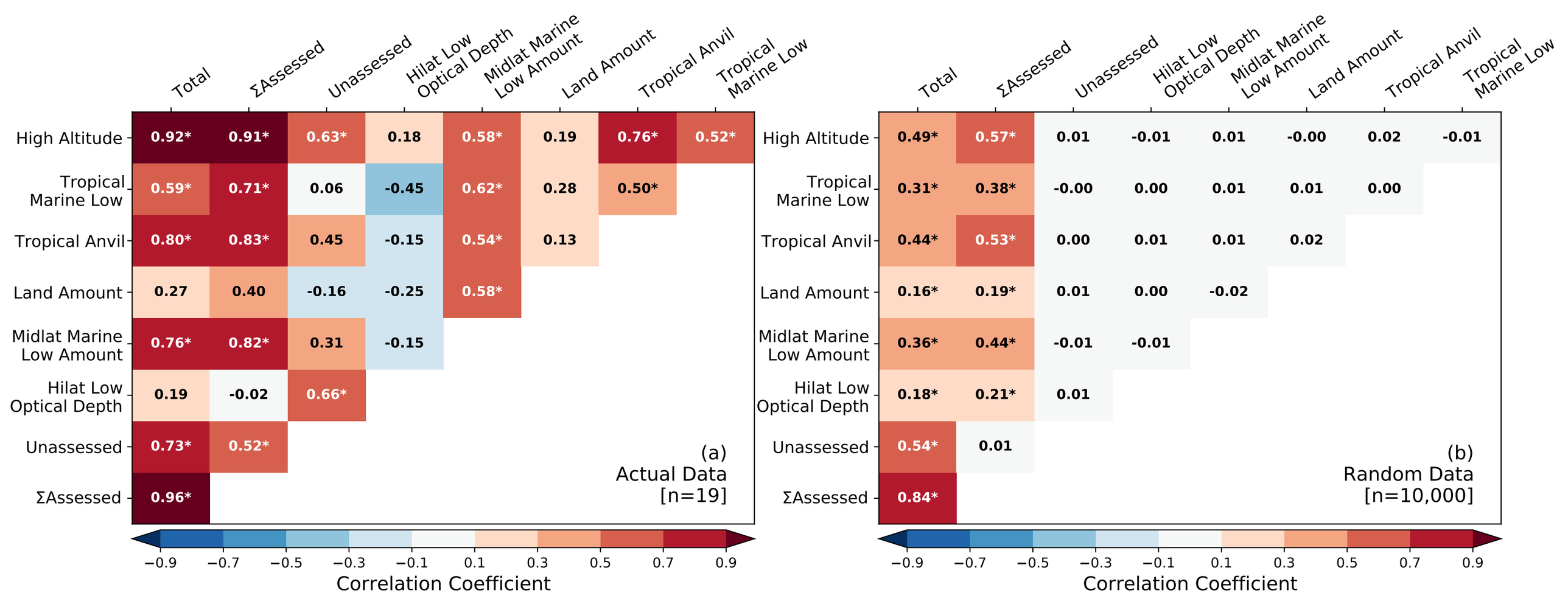
Total
ΣAssessed
Unassessed
Hilat Low Optical Depth
Midlat Marine Low Amount
Land Amount
Tropical Anvil
Tropical Marine Low
High Altitude
Tropical Marine Low
Tropical Anvil
Land Amount
Midlat Marine Low Amount
Hilat Low Optical Depth
Unassessed
ΣAssessed
0.92* 0.91* 0.63* 0.18 0.58* 0.19 0.76* 0.52*
0.59* 0.71* 0.06 -0.45 0.62* 0.28 0.50*
0.80* 0.83* 0.45 -0.15 0.54* 0.13
0.27 0.40 -0.16 -0.25 0.58*
0.76* 0.82* 0.31 -0.15
0.19 -0.02 0.66*
0.73* 0.52*
0.96*
(a)
Actual Data
[n=19]
−0.9 −0.7 −0.5 −0.3 −0.1 0.1 0.3 0.5 0.7 0.9
Correlation Coefficient
Total
ΣAssessed
Unassessed
Hilat Low Optical Depth
Midlat Marine Low Amount
Land Amount
Tropical Anvil
Tropical Marine Low
High Altitude
Tropical Marine Low
Tropical Anvil
Land Amount
Midlat Marine Low Amount
Hilat Low Optical Depth
Unassessed
ΣAssessed
0.49* 0.57* 0.01 -0.01 0.01 -0.00 0.02 -0.01
0.31* 0.38* -0.00 0.00 0.01 0.01 0.00
0.44* 0.53* 0.00 0.01 0.01 0.02
0.16* 0.19* 0.01 0.00 -0.02
0.36* 0.44* -0.01 -0.01
0.18* 0.21* 0.01
0.54* 0.01
0.84*
(b)
Random Data
[n=10,000]
−0.9 −0.7 −0.5 −0.3 −0.1 0.1 0.3 0.5 0.7 0.9
Correlation Coefficient

Figure 3.

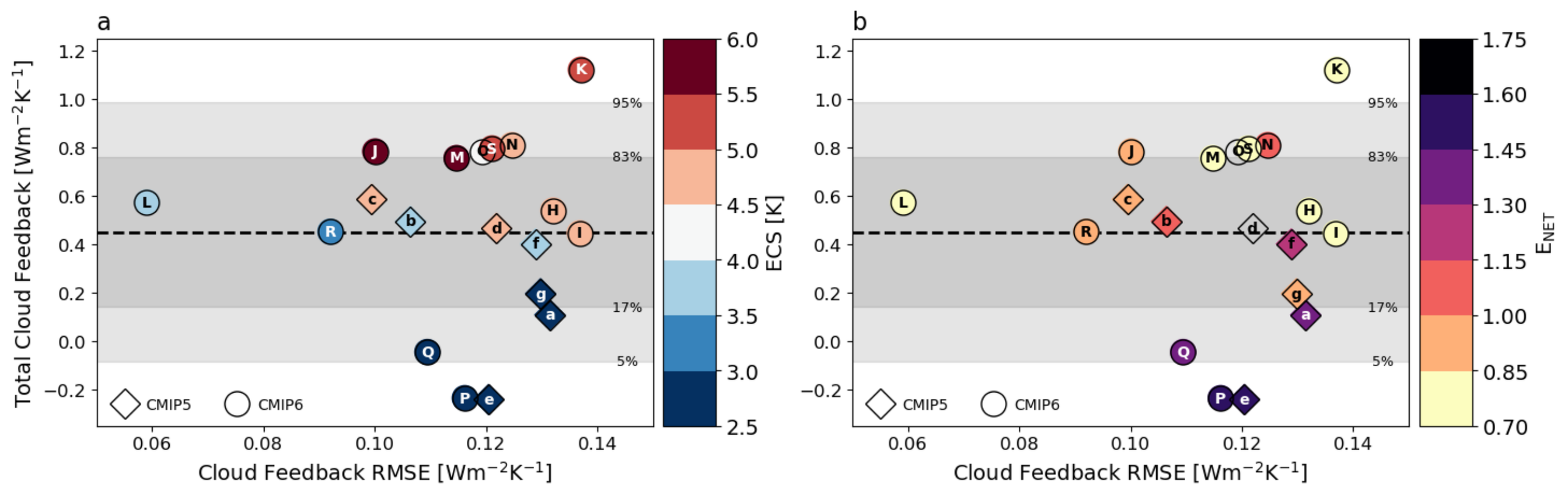
a
b
Total Cloud Feedback [Wm^{-2}K^{-1}]
Cloud Feedback RMSE [Wm^{-2}K^{-1}]
ECS [K]
E_{NET}
95%
83%
17%
5%
CMIP5
CMIP6
K
J
M
OS
N
L
c
R
b
d
H
f
I
g
a
Q
P
e

Figure 4.

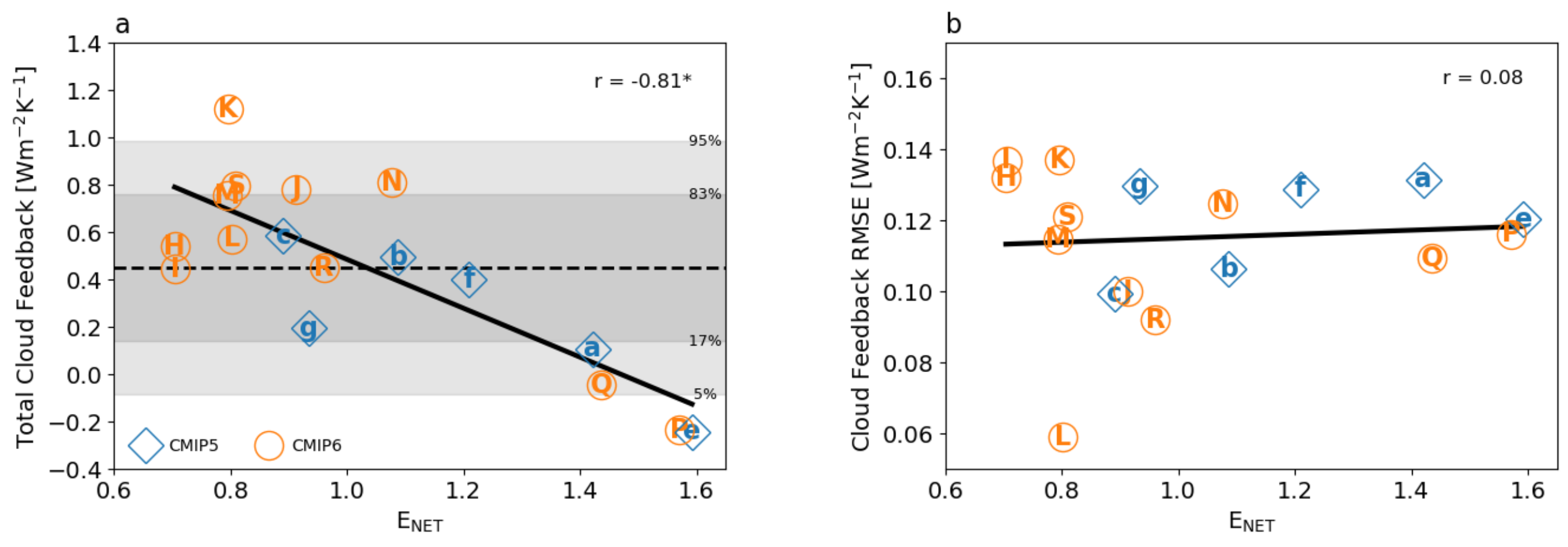
a
b
r = -0.81*
r = 0.08
95%
83%
17%
5%
CMIP5
CMIP6
Total Cloud Feedback [Wm^{-2}K^{-1}]
Cloud Feedback RMSE [Wm^{-2}K^{-1}]
E$_{NET}$
E$_{NET}$

Figure 5.

Unassessed Cloud Feedback Values [abrupt-4xCO2]

Low-Cloud Altitude
Tropical Marine Subsidence
High-Cloud Amount + Optical Depth
Tropical Land Cloud Optical Depth
60-90 Marine Low-Cloud Amount
30-40 / 70-90 Low-Cloud Optical Depth
30-90 High-Cloud Optical Depth
30-90 Marine High-Cloud Amount
Zelinka et al (2013) Decomposition Residual
Sum of Unassessed Cloud Feedbacks

CMIP5 Mean [n=7]
CMIP6 Mean [n=12]

CMIP5
CCSM4
CanESM2
HadGEM2-ES
MIROC-ESM
MIROC5
MPI-ESM-LR
MRI-CGCM3

CMIP6
CNRM-CM6-1
CNRM-ESM2-1
CanESM5
E3SM-1-0
GFDL-CM4
HadGEM3-GC31-LL
IPSL-CM6A-LR
IPSL-CM6A-LR-INCA
MIROC-ES2L
MIROC6
MRI-ESM2-0
UKESM1-0-LL

ECS [K]
4.7
3.9
2.6
2.3

−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3
$Wm^{-2}K^{-1}$