Evaluating climate models' cloud feedbacks against expert judgement

Mark D. Zelinka[1,1], Stephen A. Klein[1,1], and Yi Qin[2,2]

[1]Lawrence Livermore National Laboratory (DOE)
[2]Lawrence Livermore National Laboratory

November 30, 2022

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 Wm$ˆ{-2}$K$ˆ{-1}$) 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.

Evaluating climate models' cloud feedbacks against expert judgement

Mark D. Zelinka[1], Stephen A. Klein[1], Yi Qin[1], Timothy A. Myers[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 feedback 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 generally have moderate total cloud feedbacks (0.4–0.6 $\mathrm{Wm^{-2}K^{-1}}$) and 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 warming — 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, guarantee 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, 1990) and continuing to the most recent collection of models as part of CMIP6, the 6th phase of the Coupled Model Intercomparison Project (M. D. 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 (M. D. 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 (M. D. Zelinka et al., 2020; Flynn & Mauritsen, 2020; M. Collins et al., 2013).

Given the need for models to reliably predict future climate and the fact that cloud feedbacks strongly affect their ability to do so makes it imperative 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 expert 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 (M. D. 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 (listed in Table 1) that have successfully implemented the Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulator Package (COSP; Bodas-Salcedo et al., 2011). As will be evident below, these models exhibit cloud feedbacks spanning nearly the full range of values produced in the full ensemble of CMIP5 and CMIP6 models analyzed in M. D. Zelinka et al. (2020), and we therefore consider this subset to be a sufficiently representative sample of model diversity.

Table 1. Models used in this study. CMIP5 and CMIP6 models are indicated with lower-case and upper-case symbols, respectively. Years within the `abrupt-4xCO2` simulation with data available to analyze are indicated.

Symbol	Model	Reference	Years
a	CCSM4	Gent et al. (2011)	1-104
b	CanESM2	Arora et al. (2011)	1-21 / 121-140
c	HadGEM2-ES	W. J. Collins et al. (2011)	1-20 / 122-140
d	MIROC-ESM	S. Watanabe et al. (2011)	1-20 / 121-140
e	MIROC5	M. Watanabe and others (2010)	1-20 / 121-140
f	MPI-ESM-LR	Stevens et al. (2013)	1-20 / 121-140
g	MRI-CGCM3	Yukimoto et al. (2012)	1-20 / 121-140
H	CNRM-CM6-1	Voldoire et al. (2019)	1-150
I	CNRM-ESM2-1	Séférian et al. (2019)	1-150
J	CanESM5	Swart et al. (2019)	1-150
K	E3SM-1-0	Golaz et al. (2019)	1-150
L	GFDL-CM4	Held et al. (2019)	1-150
M	HadGEM3-GC31-LL	K. D. Williams et al. (2018)	1-150
N	IPSL-CM6A-LR	Boucher et al. (2020)	1-150
O	IPSL-CM6A-LR-INCA	Boucher et al. (2020)	1-150
P	MIROC-ES2L	Hajima et al. (2020)	1-150
Q	MIROC6	Tatebe et al. (2019)	1-150
R	MRI-ESM2-0	Yukimoto et al. (2019)	1-150
S	UKESM1-0-LL	Sellar et al. (2019)	1-150

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 (Table 1).

M. D. Zelinka et al. (2012a) validated cloud feedbacks computed using the cloud radiative kernel (CRK) methodology against independent estimates derived as the adjusted change in cloud radiative effect ($\Delta\mathrm{CRE}_{\mathrm{adj}}$; Shell et al., 2008; Soden et al., 2008) for six CMIP3 models. Here we update this comparison using the CMIP5 and CMIP6 models analyzed in this study. We compare CRK-derived cloud feedbacks with the $\Delta\mathrm{CRE}_{\mathrm{adj}}$ and approximate partial radiative perturbation (APRP; Taylor et al., 2007)-derived values computed in M. D. Zelinka et al. (2020). Six $\Delta\mathrm{CRE}_{\mathrm{adj}}$ feedbacks are provided based on the adjustments from the non-cloud radiative kernels of Soden et al. (2008), Shell et al. (2008), Block and Mauritsen (2013), Huang et al. (2017), Pendergrass et al. (2018), and Smith et al. (2018). APRP provides only the SW component, but it additionally provides estimates of SW cloud amount, scattering, and absorption feedbacks, allowing us to compare to the CRK-derived SW amount and optical depth components. Figure S1 shows the multi-model mean zonal mean SW and LW cloud feedbacks from these three techniques, along with their across-model correlations, and Figure S2 scatters the global mean CRK-derived and non-CRK-derived feedback values against each other. The CRK-derived feedbacks are in excellent agreement with the $\Delta\mathrm{CRE}_{\mathrm{adj}}$ and APRP feedbacks, for both the spatial characteristics of the multi-model mean and the across-model correlation of the zonal and global means. This confirms the validity of the CRK technique for estimating cloud feedback.

We focus in this study on feedbacks estimated from `abrupt-4xCO2` experiments so as to stay consistent with Sherwood et al. (2020), but have repeated all calculations using Atmospheric Model Intercomparison Project (`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 (with area-weighting) 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 tropical marine ascent/descent 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 M. D. 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 ≤ 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 (M. D. Zelinka et al., 2013, 2016).

Passive satellite-based measurements – like those mimicked by the ISCCP simulator used in this study – provide unobscured cloud fractions visible from space. This means that low-clouds may be hidden and revealed by changes in high-cloud cover. This complicates interpretation of low-cloud feedbacks, since high-cloud changes are aliased to an unknown extent into low-cloud feedbacks. To avoid this potential source of misinterpretation, we express the standard low-level cloud feedbacks as a sum of three terms following Scott et al. (2020) and Myers et al. (2021):

$$\text{low} = \text{low}_{\text{unobsc}} + \Delta\text{obsc} + \text{cov}.$$

$\text{low}_{\text{unobsc}}$ is the "true" low-cloud feedback occurring in regions that are not obscured by upper-level clouds and are unaffected by changes in obscuration, which we further break down into amount, altitude, optical depth, and residual components. Δobsc is the "obscuration-induced" component of low-cloud feedback arising entirely from changes in upper-level cloud fraction that reveal or hide low-level clouds. It is therefore by definition solely an "amount" component, so we absorb it into the high-cloud amount feedback. The covariance term, cov, is typically very small. To summarize, the total cloud feedback can be expressed as:

$$\text{total} = \sum_{\text{i}} \text{high}_{\text{i}} + \sum_{\text{i}} \text{low}_{\text{unobsc,i}} + \text{cov},$$

where $i \in \{\text{amount}, \text{altitude}, \text{optical depth}, \text{residual}\}$ components, and the high cloud amount component includes the Δobsc component.

In Table 2, 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. We also provide a matrix in Figure S3 to help visualize the feedback components that are computed 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. Many of the expert assessed cloud feedbacks are independent of any GCM results, but 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 expert-assessed 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 M. D. 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 over the 150-year duration of the `abrupt-4xCO2` experiment. 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.

Table 2. 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. Feedbacks are computed at each spatial location (or ω_{500} bin as appropriate), then summed over the region of interest with weighting by the fractional area of the globe represented. As explained the text, high-cloud amount feedbacks include the Δobsc term and all low-cloud feedbacks are computed using $\mathrm{low_{unobsc}}$ components.

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 high-cloud amount & optical depth	ocean	ascent	30S-30N
4. land cloud amount	0.08 ± 0.08	sum of high- and low-cloud amount	land	all	90S-90N
5. middle-latitude marine low-cloud amount	0.12 ± 0.12	low-cloud amount	ocean	all	30-60N/S
6. high-latitude low-cloud 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			

Finally, for each model we compute a radiatively-relevant cloud property error metric, E_{NET}, using Equation 5 of Klein et al. (2013). First, cloud fraction errors are computed by differencing climatological ISCCP simulator cloud fraction histograms from `amip` simulations and the ISCCP HGG observational climatology (Young et al., 2018). Both modeled 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. Second, these errors are multiplied by net (LW+SW) cloud radiative kernels, thereby weighting them by their corresponding net TOA radiative impact. Third, this product is aggregated into six cloud types: optically intermediate and thick clouds at low, middle, and high levels. These are then squared, averaged over the six categories, summed (with area weighting) over month, longitude, and latitude between 60°S and 60°N, and the square root is taken. Finally, this scalar value is normalized by the accumulated space–time standard deviation of observed radiatively-relevant cloud properties, defined analogously. This process yields a single scalar error metric, E_{NET}, in each model that quantifies the spatio-temporal error in climatological cloud properties for clouds with $\tau > 3.6$, weighted by their net TOA radiative impact. We acknowledge that evaluation against ISCCP observations is a limited viewpoint on the quality of models' cloud simulations — one that may change if using other cloud datasets, like those derived from active sensors.

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 in Table 2. 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 3, 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 4-22 are identical to Figure 1, but with individual models highlighted in each figure for better discrimination.

All but seven models fall within the *likely* range assessed for the high cloud altitude feedback and the multi-model means are 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 wide inter-model variation is noteworthy 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. Only one model (CanESM5) exceeds the central expert assessed value, and several models' values lie below the lower bound of the *likely* (MIROC5, MRI-CGCM3, HadGEM3-GC31-LL, MIROC-ES2L, and MIROC6) and *very likely* (CCSM4) confidence intervals.

In contrast, all models underestimate the strength of the negative anvil cloud feedback, relative to the central value assessed in Sherwood et al. (2020). Eight models (MRI-CGCM3, CNRM-CM6-1, CNRM-ESM2-1, E3SM-1-0, HadGEM3-GC31-LL, IPSL-CM6A-

Table 3. 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	-0.05**	-0.07	0.08	0.07	-0.03	-0.01	0.12*	0.11*	0.14	1.42	2.94
b) CanESM2	r1i1p1	0.27	0.14	-0.06	0.06	0.05	-0.05	0.07	0.42	0.49	0.09	1.09	3.70
c) HadGEM2-ES	r1i1p1	0.30*	0.15	-0.02	0.07	0.11	-0.06	0.03	0.56	0.59	0.10	0.89	4.60*
d) MIROC-ESM	r1i1p1	0.17	0.15	-0.04	0.10	0.13	-0.14*	0.09	0.38	0.47	0.10	N/A	4.65*
e) MIROC5	r1i1p1	0.00**	0.07*	-0.14	0.03	-0.04*	-0.05	-0.13	-0.11**	-0.24**	0.13	1.59	2.71
f) MPI-ESM-LR	r1i1p1	0.17	0.22	-0.05	0.08	0.11	-0.16*	0.03	0.37	0.40	0.09	1.21	3.63
g) MRI-CGCM3	r1i1p1	0.10*	0.07*	0.03*	0.03	0.04	-0.06	-0.02	0.22	0.20	0.13	0.93	2.61
CMIP5 Average		0.16	0.11	-0.05	0.07	0.07	-0.08	0.01	0.28	0.29	0.11	1.19	3.55
CMIP5 1-σ		0.10	0.08	0.05	0.02	0.05	0.05	0.07	0.20	0.27	0.02	0.25	0.78
H) CNRM-CM6-1	r1i1p1f2	0.27	0.06*	0.02*	0.04	0.05	-0.01	0.12	0.42	0.54	0.13	0.70	4.90**
I) CNRM-ESM2-1	r1i1p1f2	0.23	0.04*	0.02*	0.03	0.02	-0.01	0.11	0.34	0.45	0.13	0.71	4.79**
J) CanESM5	r1i1p2f1	0.30*	0.27	-0.06	0.05	0.09	-0.03	0.17	0.62	0.78*	0.08	0.91	5.62**
K) E3SM-1-0	r1i1p1f1	0.38**	0.21	0.01*	0.09	0.21	-0.02	0.24	0.88*	1.12**	0.12	0.80	5.31**
L) GFDL-CM4	r1i1p1f1	0.19	0.17	-0.12	0.09	0.19	-0.05	0.11	0.46	0.57	0.06	0.80	3.89
M) HadGEM3-GC31-LL	r1i1p1f3	0.20	0.09*	0.03*	0.07	0.25*	-0.01	0.12	0.64	0.76	0.12	0.79	5.55**
N) IPSL-CM6A-LR	r1i1p1f1	0.29	0.13	0.02*	0.13	0.21	-0.04	0.05	0.76	0.81*	0.12	1.08	4.70*
O) IPSL-CM6A-LR-INCA	r1i1p1f1	0.27	0.13	0.02*	0.14	0.21	-0.04	0.05	0.73	0.78*	0.12	N/A	4.13*
P) MIROC-ES2L	r1i1p1f2	0.01**	0.06*	-0.19	0.05	-0.01*	-0.03	-0.11	-0.12**	-0.23**	0.12	1.57	2.66
Q) MIROC6	r1i1p1f1	0.09*	0.05*	-0.08	0.10	-0.05*	-0.04	-0.11	0.06*	-0.05*	0.13	1.44	2.60
R) MRI-ESM2-0	r1i1p1f1	0.24	0.15	-0.06	0.01	0.12	-0.04	0.03	0.43	0.45	0.08	0.96	3.13
S) UKESM1-0-LL	r1i1p1f2	0.23	0.10	0.02*	0.06	0.25*	-0.02	0.15	0.65	0.80*	0.12	0.81	5.36**
CMIP6 Average		0.23	0.12	-0.03	0.07	0.13	-0.03	0.08	0.49	0.56	0.11	0.96	4.39
CMIP6 1-σ		0.09	0.07	0.07	0.04	0.10	0.01	0.10	0.28	0.36	0.02	0.28	1.05
CMIP5/6 Average		0.20	0.12	-0.04	0.07	0.11	-0.05	0.05	0.41	0.46	0.11	1.04	4.08
CMIP5/6 1-σ		0.10	0.07	0.06	0.03	0.09	0.04	0.10	0.27	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			

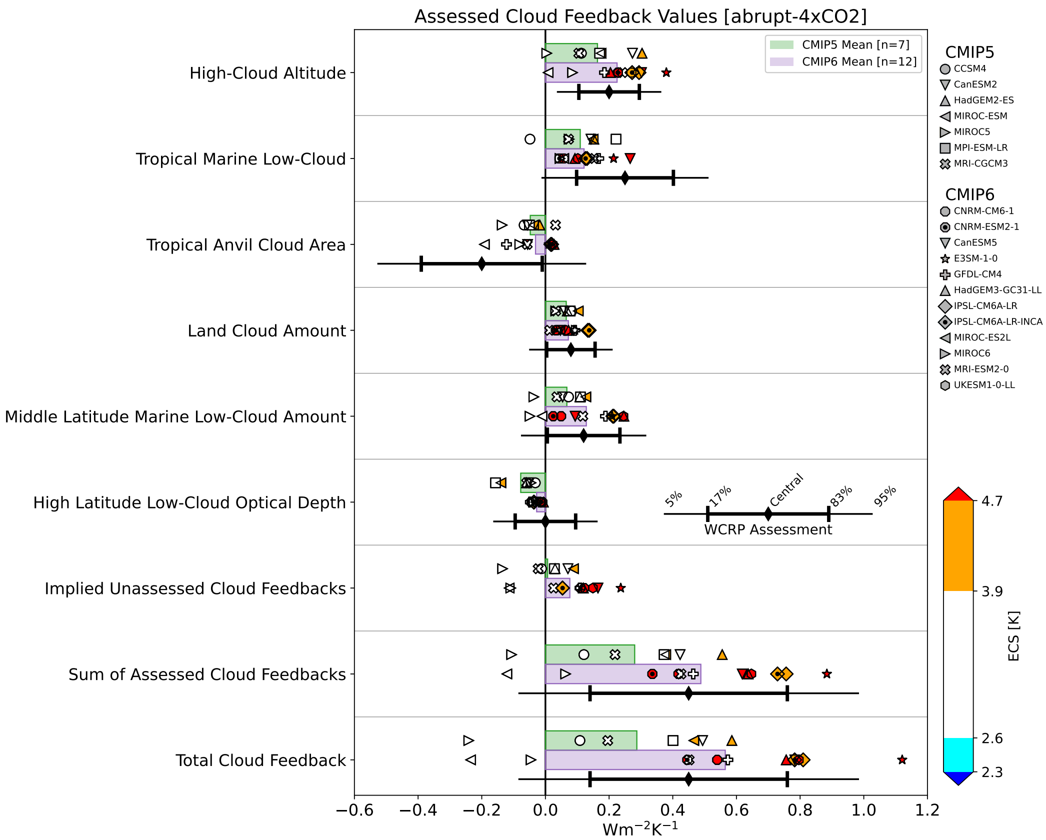

Figure 1. Cloud feedback components estimated from climate model simulations and as assessed in Sherwood et al. (2020). For each component, the individual model values are indicated with symbols, the multi-model means are indicated with green (CMIP5) and purple (CMIP6) bars, and the expert assessed *likely* 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 S4-S22.

LR, IPSL-CM6A-LR-INCA, and UKESM1-0-LL) have positive anvil feedbacks that place them above the upper bound of the assessed *likely* confidence interval.

All models lie within the assessed *likely* range for the land cloud amount feedback, while all but five models (MIROC5, HadGEM3-GC31-LL, MIROC-ES2L, MIROC6, and UKESM1-0-LL) lie within the assessed *likely* range of the middle latitude marine low cloud amount feedback.

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; M. D. Zelinka et al., 2020; Flynn & Mauritsen, 2020). The inter-model spread in this feedback component has also dramatically decreased in CMIP6.

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 larger than all other individual components except the high cloud altitude feedback. 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 ($\sigma = 0.27$ 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 = 0.20 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 E3SM-1-0 lies above the upper bound of the assessed *likely* confidence interval.

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 (MIROC-ESM, MPI-ESM-LR, CNRM-ESM2-1, and MRI-ESM2-0) 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 now four new models (CanESM5, IPSL-CM6A-LR, IPSL-CM6A-LR-INCA, and UKESM1-0-LL) that 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 five models with total cloud feedbacks above the upper limit of the expert-assessed *likely* range (CanESM5, E3SM-1-0, IPSL-CM6A-LR, IPSL-CM6A-LR-INCA, and UKESM1-0-LL) 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 all but IPSL-CM6A-LR and IPSL-CM6A-LR-INCA 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 these models to very high ECS. No models considered here — even those whose cloud feedbacks lie below the lower limit of the *likely* and *very likely* total cloud feedback confidence bound — have ECS values below 2.6 K, the lower limit of the Sherwood et al. (2020) assessed *likely* range. In general, too-large cloud feedbacks seem to guarantee too-large ECS, but too-small cloud feedbacks do not guarantee too-small ECS. Also, too-large ECS can arise even without too-large cloud feedbacks.

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 M. D. 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 change very little on average.

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 (-0.04 ± 0.06 $\mathrm{Wm^{-2}K^{-1}}$), 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, midlatitude marine low cloud amount, and unassessed feedbacks (Figure 2a, column 1). 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, but it also may be related to the precise definition of low-cloud types: Taking the sum of stratocumulus and trade cumulus cloud feedbacks diagnosed in Myers et al. (2021) using different meteorological criteria than employed here as an alternative estimate of tropical marine low-cloud feedback, we find a larger correlation (r=0.80) with total cloud feedback.

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

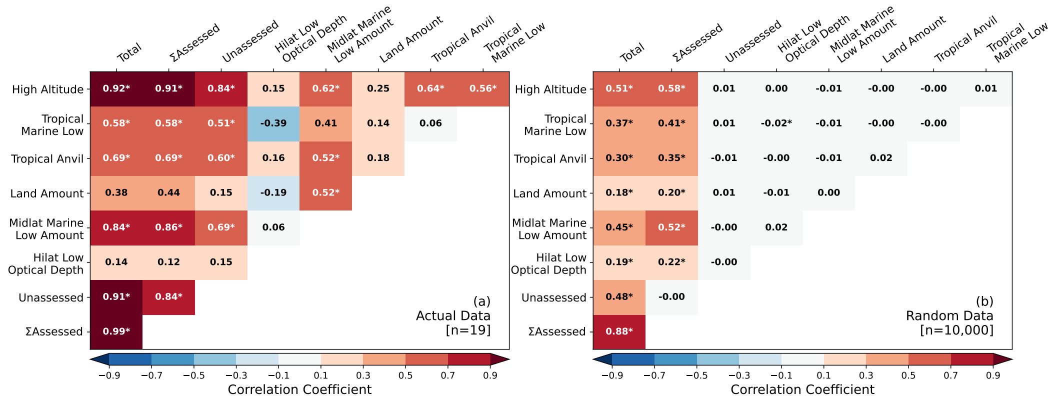

Figure 2. Matrix showing the across-model correlation among all cloud feedback components for (a) actual model data and (b) synthetic normally-distributed data with means and standard deviations equal to those of the models for each feedback component. Correlations that are significantly different from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

cloud feedback and (2) a wider spread in the total cloud feedback than would occur if 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 largest values in all categories except the land cloud feedback (Figure S14 and Table 3). 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 S8 and Table 3). 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 CNRM-ESM2-1, which has feedbacks on the high tail of the model distribution for some components and on the low tail for others (Figure S12 and Table 3).

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. As will be shown in Section 3.5, high-cloud feedbacks are among the largest components of the unassessed feedback; hence it is plausible that the positive correlations among the unassessed, high-cloud altitude, and anvil feedbacks reflect a shared physical mechanism involving high clouds. 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 3.4.

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 3 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 values are the same for both ensembles (Table

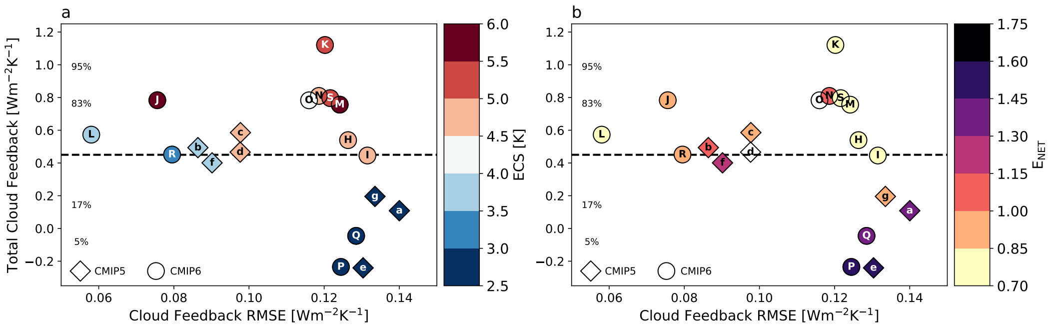

Figure 3. Total cloud feedback scattered against cloud feedback RMSE, with expert *likely* and *very likely* ranges of total cloud feedback indicated with horizontal shading. Models are denoted by the symbols listed in Table 3 and are colored according to their (a) ECS values and (b) net radiatively-relevant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

3). Although the three 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, the same number of models show degraded performance as improved performance in this measure: MIROC-ES2L [P] and the two UKMO models (HadGEM3-GC31-LL [M] and UKESM1-0-LL [S]) have higher RMSE than their predecessors (MIROC-ESM [d], and HadGEM2-ES [c]), whereas CanESM5 [J], MIROC6 [Q], and MRI-ESM2-0 [R] have lower RMSE than their predecessors (CanESM2 [b], MIROC5 [e], and MRI-CGCM3 [g]).

The seven models with smaller-than-average cloud feedback errors (i.e., RMSE $\leq$ 0.11 $\mathrm{Wm^{-2}K^{-1}}$) have moderate (0.4–0.6 $\mathrm{Wm^{-2}K^{-1}}$) total cloud feedbacks, except for CanESM5 [J], which has a total cloud feedback of 0.8 $\mathrm{Wm^{-2}K^{-1}}$. All but three of these models have moderate (3–4 K) ECS values, the exceptions being HadGEM2-ES [c], MIROC-ESM [d], and CanESM5 [J], which have ECS values above 4.5 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). Of the seven models with below-average feedback errors, GFDL-CM4 [L], MRI-ESM2-0 [R], and CanESM2 [b] are the only ones for which all assessed feedbacks lie within the expert *likely* range (Figures S15, S21, and S5, respectively; Table 3). 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. All five models with small total cloud feedback ($<$ 0.2 $\mathrm{Wm^{-2}K^{-1}}$) and small ECS ($<$ 3 K) have cloud feedback components that are systematically biased low relative to expert judgement, giving them larger-than-average RMSE. Most 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], MIROC-ESM [d], and CanESM5 [J] have below-average RMSE value. CCSM4 [a] has the highest RMSE of all models considered despite lying within the assessed *likely* range for five components (Figure S4; Table 3).

Two models (CNRM-CM6-1 [H] and CNRM-ESM2-1 [I]) have total cloud feedbacks very close to the central value of the expert assessment but larger-than-average RMSE values. They achieve reasonable total cloud feedbacks partly through having low-biased

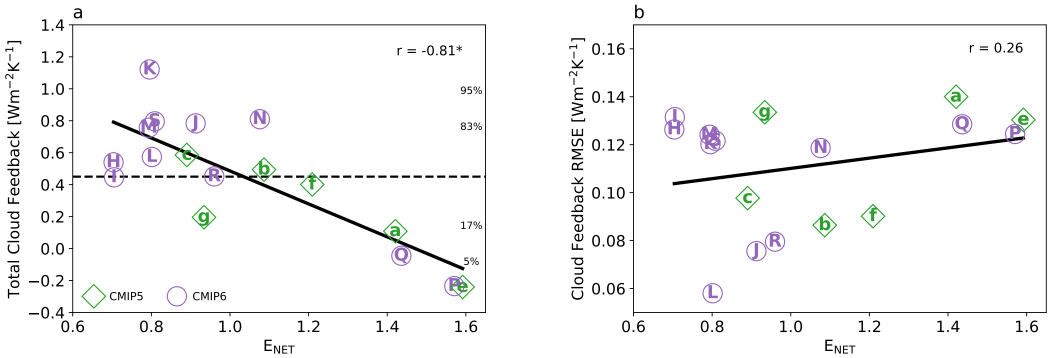

Figure 4. (a) Total cloud feedback and (b) cloud feedback RMSE scattered against net radiatively-relevant cloud property error metric, E_{NET}. Models are denoted by the symbols listed in Table 3 and are colored green for CMIP5 and purple for CMIP6. Expert *likely* and *very likely* ranges of total cloud feedback indicated with horizontal shading in (a). Correlations that are significant at 95% confidence are indicated with an asterisk.

tropical marine low cloud feedbacks that counteract their high-biased tropical anvil cloud area feedbacks (Figures S11-12; Table 3). Put simply, they get the right answer for the wrong reasons.

GFDL-CM4, CanESM5, MRI-ESM2-0, and CanESM2 remain the four 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 (E_{NET}), 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). Given that E_{NET} is an aggregated metric, we also tested whether the anticorrelation persists when considering relationships between individual cloud feedbacks and cloud-type specific E_{NET} values (e.g., between midlatitude marine low-cloud amount feedback and mean-state errors for midlatitude marine low-clouds). This anticorrelation continues to hold for all but the land cloud amount feedback, albeit with weaker correlation coefficients (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 3). 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 (encapsulated in RMSE) to the agreement with observations of the present-day climatological distribution of clouds and their properties (encapsulated in E_{NET}). While the correlation between these two metrics is positive, it is very weak and not significant at 95% confidence. Moreover, many models with small mean-state cloud errors have cloud feedback errors that are as large or larger than models with large mean-state errors, indicating that improved simulation of mean-state cloud properties does not necessarily lead to improved cloud feedbacks with respect to expert judgment. The weak correlation also holds for relationships between RMSE and components of E_{NET} corresponding to individual cloud feedbacks (not shown).

In Figure 3b, 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 errors 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.

While realistic mean-state cloud properties may not guarantee that a model simulates more reliable cloud feedbacks, the models with worst mean-state cloud properties (i.e., $E_{NET} > 1.3$) all have poor agreement with the expert-assessed total cloud feedback and/or its components (see models at top right of Figure 4b). This is also evidenced by the fact that most of the models with large mean-state errors (purple/black shading) have large 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). This suggests that simulating poor mean-state cloud properties precludes a model from simulating cloud feedbacks in agreement with expert judgement. In other words, better simulation of mean-state cloud properties may be a necessary but insufficient criterion for simulating more trustworthy cloud feedbacks.

This finding has support in recent literature. Mülmenstädt et al. (2021) showed that a model with better mean-state cloud properties could have greater biases in its climate responses owing to compensating errors in cloud and precipitation processes. As noted in that study, fidelity in simulating mean-state clouds alone is an insufficient constraint on a model's feedback because of the many different combinations of process representations that can lead to equally valid representations of mean-state clouds. Since these process representations can all differ in their sensitivity to warming, the cloud feedback is not uniquely determined by mean-state properties, and improving the representation of the mean-state (especially at the expense of the process-level) does not guarantee that feedbacks will be more reliably simulated. This notion is supported by the fact that the set of model parameters driving the variance in mean-state extratropical cloud radiative effect across members of the HadGEM3-GA7.05 perturbed physics ensemble differ from those driving the variance in its cloud feedback (Tsushima et al., 2020). A corollary to this are the many examples in which models with better "bottom-up" process representation more poorly satisfy "top-down" constraints like the observed historical global mean temperature evolution (Golaz et al., 2013; Suzuki et al., 2013), expert-assessed magnitude of aerosol indirect effects (Jing & Suzuki, 2018) or paleoclimate states (Zhu et al., 2020, 2021)

3.5 GCM Cloud Feedbacks in Unassessed Categories

Sherwood et al. (2020) only assessed quantitative values for a selection of well-studied cloud feedbacks, so it is important to know whether any of the unassessed feedbacks are substantial. 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. Figure 5 shows a breakdown of explicitly-computed feedbacks that were not assessed in Sherwood et al. (2020). 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). See Figure S3 for a matrix that helps to visualize and rationalize the discretization made.

The multi-model mean unassessed cloud feedback transitions from being 0.01 $Wm^{-2}K^{-1}$ on average in CMIP5 to 0.08 $Wm^{-2}K^{-1}$ on average in CMIP6. The largest shift occurs for the multi-model mean extratropical high cloud optical depth component, which transitions from a negative to a weak positive value. This component, along with the tropical marine ascent low-cloud amount plus optical depth component exhibit the largest inter-model spread among all unassessed categories, and may be worthwhile targets for future expert assessment.

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. MIROC5, MIROC-ES2L, and MIROC6 exhibit strong negative unassessed cloud feedbacks (with values < -0.10 $Wm^{-2}K^{-1}$) that are comparable in magnitude to the sum of their assessed feedbacks. MIROC5 and MIROC6 have strong negative low-cloud amount plus optical depth components in tropical marine ascent regions, while MIROC-ES2L has strong negative high-cloud amount and optical depth components in tropical marine subsidence regions. All three of these models have moderately negative extratropical high-cloud optical depth feedbacks as well. Two CMIP6 models (CanESM5 and E3SM-1-0) have positive unassessed feedbacks that exceed 0.15 $Wm^{-2}K^{-1}$ — the multi-model mean plus standard deviation. This occurs because of several systematically positive components, the largest of which is the 0.11 $Wm^{-2}K^{-1}$ extratropical high-cloud optical depth component in E3SM-1-0.

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, in general, models that most closely match the expert-assessed values across several cloud feedback components have moderate total cloud feedbacks (0.4–0.6 $Wm^{-2}K^{-1}$) and 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 three best models in our error metric are CMIP6 models, 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*

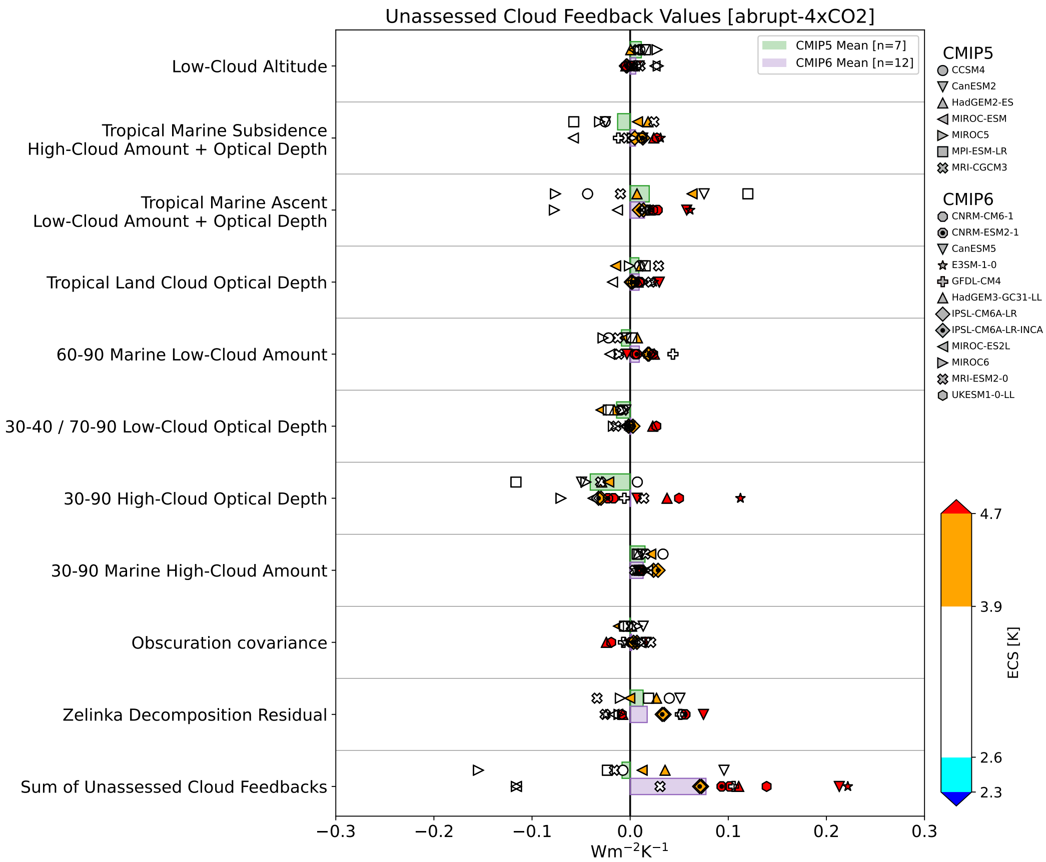

Figure 5. As in Figure 1, but for cloud feedback components that were not assessed in Sherwood et al. (2020). Note the x-axis spans a range that is only a third of that in Figure 1.

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 2xCO_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. The reasons for this remain to be investigated, but it is consistent with emergent constraint studies involving mean-state properties of clouds or their environment, nearly all of which point to higher-than-average cloud feedbacks and climate sensitivities (Volodin, 2008; Trenberth & Fasullo, 2010; Fasullo & Trenberth, 2012; Sherwood et al., 2014; Tian, 2015; Brient et al., 2016; Siler et al., 2018).

But more skillful simulation of mean-state cloud properties does not guarantee 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. Skill at simulating mean-state cloud properties appears to be a necessary but not sufficient criterion for simulating realistic cloud feedbacks.

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. Establishing and understanding the physical basis of correlations among feedback components and their potential linkages with mean-state cloud properties is important future work.

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 (also seen in Myers et al., 2021), and may be tied to reductions in super-cooled liquid biases in the latest models (Bodas-Salcedo et al., 2019; Gettelman et al., 2019; M. D. Zelinka et al., 2020).

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; Ceppi & Nowack, 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, eight 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 and Stevens (2015) and I. N. Williams and 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 partly drive the increase in multi-model average cloud feedback from CMIP5 to CMIP6. Of these, the extratropical high cloud optical depth component exhibits the largest increase. 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 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). So doing could help modelers to identify and correct erroneous cloud feedbacks that lead to biased climate sensitivity prior to the model being frozen, thereby increasing the reliability of the model for policy-relevant climate projections (e.g., Voosen, 2021).

Acknowledgments

Python code to perform all calculations and produce all figures and tables in this manuscript is available at `https://doi.org/10.5281/zenodo.5206838` (M. Zelinka, 2021a) and is being incorporated into the PCMDI Metrics Package (Doutriaux et al., 2018), available at `https://doi.org/10.5281/zenodo.1414560`. CMIP5 and CMIP6 ECS values are available at `https://doi.org/10.5281/zenodo.5206851` (M. Zelinka, 2021b). 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 DE-AC52-07NA27344. We are grateful for stimulating discussions with Leo Donner, Chris Golaz, Yoko Tsushima, and Mark Webb, and for the helpful comments from three anonymous reviewers.

References

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., ... Merryfield, W. J. (2011). Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophys. Res. Lett.*, *38*. doi: 10.1029/2010GL046270

Block, K., & Mauritsen, T. (2013). Forcing and feedback in the MPI-ESM-LR coupled model under abruptly quadrupled CO2. *Journal of Advances in Modeling Earth Systems*, *5*(4), 676–691. doi: 10.1002/jame.20041

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 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

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., ... Vuichard, N. (2020). Presentation and Evaluation of the IPSL-CM6A-LR Climate Model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. Retrieved 2021-07-17, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002010` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS002010) doi: 10.1029/2019MS002010

Brient, F., Schneider, T., Tan, Z., Bony, S., Qu, X., & Hall, A. (2016, July). Shallowness of tropical low clouds as a predictor of climate models' response to warming. *Climate Dynamics*, *47*(1), 433–449. Retrieved 2021-08-04, from `https://doi.org/10.1007/s00382-015-2846-0` doi: 10.1007/s00382-015-2846-0

Ceppi, P., & Nowack, P. (2021, July). Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, *118*(30). Retrieved 2021-08-09, from `https://www.pnas.org/content/118/30/e2026290118` (Publisher: National Academy of Sciences Section: Physical Sciences) doi: 10.1073/pnas.2026290118

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., Potter, G. L., Blanchet, J. P., Boer, G. J., Genio, A. D. D., Déqué, M., … Zhang, M.-H. (1990). Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *Journal of Geophysical Research: Atmospheres*, *95*(D10), 16601–16615. Retrieved 2021-08-10, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JD095iD10p16601` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/JD095iD10p16601) doi: 10.1029/JD095iD10p16601

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.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Hinton, P. H. T., Hughes, J., … Woodward, S. (2011). Development and evaluation of an Earth-system model - HadGEM2. *Geosci. Model Dev. Discuss.*, *4*, 997–1062.

Doutriaux, C., Gleckler, P., Durack, P. J., Lee, J., Covey, C., Sperber, K., … jservonnat (2018, September). *PCMDI/pcmdi_metrics: PMP Version 1.2.* Zenodo. Retrieved 2021-08-16, from `https://zenodo.org/record/1414560` doi: 10.5281/zenodo.1414560

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

Fasullo, J. T., & Trenberth, K. E. (2012). A Less Cloudy Future: The Role of Subtropical Subsidence in Climate Sensitivity. *Science*, *338*(6108), 792–794. doi: 10.1126/science.1227465

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., & Dütsch, 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

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., … Zhang, M. (2011). The community climate system model version 4. *J. Climate*, *24*, 4973–4991. doi: 10.1175/2011JCLI4083.1

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

Golaz, J. C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., … Zhu, Q. (2019). The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2089–2129. doi: 10.1029/2018ms001603

Golaz, J.-C., Horowitz, L. W., & Levy, H. (2013). Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, *40*(10), 2246–2251. Retrieved 2021-07-24, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50232 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/grl.50232) doi: 10.1002/grl.50232

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., ... Kawamiya, M. (2020, May). Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geoscientific Model Development*, *13*(5), 2197–2244. Retrieved 2021-07-17, from https://gmd.copernicus.org/articles/13/2197/2020/ (Publisher: Copernicus GmbH) doi: 10.5194/gmd-13-2197-2020

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

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., ... Zadeh, N. (2019). Structure and Performance of GFDL's CM4.0 Climate Model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691–3727. Retrieved 2021-07-17, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001829 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001829) doi: 10.1029/2019MS001829

Huang, Y., Xia, Y., & Tan, X. X. (2017). On the pattern of CO2 radiative forcing and poleward energy transport. *Journal of Geophysical Research-Atmospheres*, *122*(20), 10578–10593. doi: 10.1002/2017jd027221

Jing, X., & Suzuki, K. (2018). The Impact of Process-Based Warm Rain Constraints on the Aerosol Indirect Effect. *Geophysical Research Letters*, *45*(19), 10,729–10,737. Retrieved 2021-07-26, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL079956 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL079956) doi: 10.1029/2018GL079956

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

Myers, T. A., Scott, R. C., Zelinka, M. D., Klein, S. A., Norris, J. R., & Caldwell, P. M. (2021, June). Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. *Nature Climate Change*, *11*(6), 501–507. Retrieved 2021-07-12, from https://www.nature.com/articles/s41558-021-01039-0 (Number: 6 Publisher: Nature Publishing Group) doi: 10.1038/s41558-021-01039-0

Mülmenstädt, J., Salzmann, M., Kay, J. E., Zelinka, M. D., Ma, P.-L., Nam, C., ... Quaas, J. (2021, June). An underestimated negative cloud feedback from cloud lifetime changes. *Nature Climate Change*, *11*(6), 508–513. Retrieved 2021-07-

24, from https://www.nature.com/articles/s41558-021-01038-1 (Number: 6 Publisher: Nature Publishing Group) doi: 10.1038/s41558-021-01038-1

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

Pendergrass, A. G., Conley, A., & Vitt, F. M. (2018). Surface and top-of-atmosphere radiative feedback kernels for CESM-CAM5. *Earth System Science Data*, *10*(1), 317–324. doi: 10.5194/essd-10-317-2018

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

Scott, R. C., Myers, T. A., Norris, J. R., Zelinka, M. D., Klein, S. A., Sun, M., & Doelling, D. R. (2020, September). Observed Sensitivity of Low-Cloud Radiative Effects to Meteorological Perturbations over the Global Oceans. *Journal of Climate*, *33*(18), 7717–7734. Retrieved 2021-04-21, from https://journals.ametsoc.org/view/journals/clim/33/18/jcliD191028.xml (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI-D-19-1028.1

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

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., ... Zerroukat, M. (2019). UKESM1: Description and Evaluation of the U.K. Earth System Model. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4513–4558. Retrieved 2021-07-17, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001739 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001739) doi: 10.1029/2019MS001739

Shell, K. M., Kiehl, J. T., & Shields, C. A. (2008). Using the Radiative Kernel Technique to Calculate Climate Feedbacks in NCAR's Community Atmospheric Model. *J. Climate*, *21*(10), 2269–2282. doi: 10.1175/2007JCLI2044.1

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37–42. doi: 10.1038/nature12829

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

Siler, N., Po-Chedley, S., & Bretherton, C. S. (2018). Variability in modeled cloud feedback tied to differences in the climatological spatial pattern of clouds. *Climate Dynamics*, *50*(3), 1209–1220. doi: 10.1007/s00382-017-3673-2

Smith, C. J., Kramer, R. J., Myhre, G., Forster, P. M., Soden, B. J., Andrews, T., ... Watson-Parris, D. (2018). Understanding Rapid Adjustments to Diverse

Forcing Agents. *Geophysical Research Letters*, *45*(21), 12023–12031. doi: 10.1029/2018gl079826

Soden, B. J., Held, I. M., Colman, R., Shell, K. M., Kiehl, J. T., & Shields, C. A. (2008). Quantifying Climate Feedbacks Using Radiative Kernels. *J. Climate*, *21*, 3504–3520. doi: 10.1175/2007JCLI2110.1

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., ... Roeckner, E. (2013). Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. doi: 10.1002/jame.20015

Suzuki, K., Golaz, J.-C., & Stephens, G. L. (2013). Evaluating cloud tuning in a climate model with satellite observations. *Geophysical Research Letters*, *40*(16), 4464–4468. Retrieved 2021-07-26, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50874` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/grl.50874) doi: 10.1002/grl.50874

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., ... Winter, B. (2019, November). The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*(11), 4823–4873. Retrieved 2021-07-17, from `https://gmd.copernicus.org/articles/12/4823/2019/` (Publisher: Copernicus GmbH) doi: 10.5194/gmd-12-4823-2019

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., ... Madec, G. (2019). Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4182–4227. Retrieved 2021-07-17, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001791` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001791) doi: 10.1029/2019MS001791

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., ... Kimoto, M. (2019, July). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765. Retrieved 2021-07-17, from `https://gmd.copernicus.org/articles/12/2727/2019/` (Publisher: Copernicus GmbH) doi: 10.5194/gmd-12-2727-2019

Taylor, K. E., Crucifix, M., Braconnot, P., Hewitt, C. D., Doutriaux, C., Broccoli, A. J., ... Webb, M. J. (2007). Estimating Shortwave Radiative Forcing and Response in Climate Models. *J. Climate*, *20*(11), 2530–2543. doi: 10.1175/JCLI4143.1

Tian, B. (2015). Spread of model climate sensitivity linked to double-Intertropical Convergence Zone bias. *Geophysical Research Letters*, *42*(10), 4133–4141. Retrieved 2021-08-04, from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL064119` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015GL064119) doi: 10.1002/2015GL064119

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

Trenberth, K. E., & Fasullo, J. T. (2010). Simulation of Present-Day and Twenty-First-Century Energy Budgets of the Southern Oceans. *J. Climate*, *23*(2), 440–454. doi: 10.1175/2009JCLI3152.1

Tsushima, Y., Ringer, M. A., Martin, G. M., Rostron, J. W., & Sexton, D. M. H.

(2020, September). Investigating physical constraints on climate feedbacks using a perturbed parameter ensemble. *Climate Dynamics*, *55*(5), 1159–1185. Retrieved 2021-07-24, from https://doi.org/10.1007/s00382-020-05318-y doi: 10.1007/s00382-020-05318-y

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., … Waldman, R. (2019). Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213. Retrieved 2021-07-17, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001683 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001683) doi: 10.1029/2019MS001683

Volodin, E. M. (2008). Relation between temperature sensitivity to doubled carbon dioxide and the distribution of clouds in current climate models. *Izvestiya, Atmospheric and Oceanic Physics*, *44*(3). Retrieved 2021-08-04, from https://link.springer.com/epdf/10.1134/s0001433808030043 doi: 10.1134/s0001433808030043

Voosen, P. (2021, July). *U.N. climate panel confronts implausibly hot forecasts of future warming.* Retrieved 2021-07-28, from https://www.sciencemag.org/news/2021/07/un-climate-panel-confronts-implausibly-hot-forecasts-future-warming

Watanabe, M., & others. (2010). Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *J. Climate*, *23*, 6312–6335.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., … Kawamiya, M. (2011). MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, *4*(4), 845–872. doi: 10.5194/gmd-4-845-2011

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

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., … Xavier, P. K. (2018). The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations. *Journal of Advances in Modeling Earth Systems*, *10*(2), 357–380. Retrieved 2021-07-17, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS001115 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017MS001115) doi: 10.1002/2017MS001115

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

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., ... Kitoh, A. (2012). A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3 —Model Description and Basic Performance—. *Journal of the Meteorological Society of Japan. Ser. II*, *90A*, 23–64. doi: 10.2151/jmsj.2012-A02

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., ... Ishii, M. (2019). The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component. *Journal of the Meteorological Society of Japan. Ser. II*, *97*(5), 931–965. doi: 10.2151/jmsj.2019-051

Zelinka, M. (2021a, August). *mzelinka/assessed-cloud-fbks: Aug 16, 2021 Release.* Zenodo. Retrieved 2021-08-16, from https://zenodo.org/record/5206838 doi: 10.5281/zenodo.5206838

Zelinka, M. (2021b, August). *mzelinka/cmip56_forcing_feedback_ecs: Aug 16, 2021 Release.* Zenodo. Retrieved 2021-08-16, from https://zenodo.org/record/5206851 doi: 10.5281/zenodo.5206851

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., Po-Chedley, 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., Otto-Bliesner, 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

Supporting Information for "Evaluating climate models' cloud feedbacks against expert judgement"

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

[1]Lawrence Livermore National Laboratory

Contents of this file

1. Figures S1 to S22

Introduction

In this document, we provide 22 supplementary figures. Figures S1 and S2 compare cloud radiative kernel-derived cloud feedbacks with those derived using independent methods. Figure S3 provides a matrix showing which regions and cloud types contribute to each feedback, facilitating understanding of how the assessed feedbacks are computed, which feedbacks are left unassessed, and how we further discretize these remaining unassessed feedbacks. Figures S4-S22 are identical to Figure 1 of the main text, but individual models are highlighted in each.

References

Huang, Y., Xia, Y., & Tan, X. X. (2017). On the pattern of CO2 radiative forcing and poleward energy transport. *Journal of Geophysical Research-Atmospheres*, *122*(20), 10578–10593. doi: 10.1002/2017jd027221

Shell, K. M., Kiehl, J. T., & Shields, C. A. (2008). Using the Radiative Kernel Technique to Calculate Climate Feedbacks in NCAR's Community Atmospheric Model. *J. Climate*, *21*(10), 2269–2282. doi: 10.1175/2007JCLI2044.1

Soden, B. J., Held, I. M., Colman, R., Shell, K. M., Kiehl, J. T., & Shields, C. A. (2008). Quantifying Climate Feedbacks Using Radiative Kernels. *J. Climate*, *21*, 3504–3520. doi: 10.1175/2007JCLI2110.1

Taylor, K. E., Crucifix, M., Braconnot, P., Hewitt, C. D., Doutriaux, C., Broccoli, A. J., ... Webb, M. J. (2007). Estimating Shortwave Radiative Forcing and Response in Climate Models. *J. Climate*, *20*(11), 2530–2543. doi: 10.1175/JCLI4143.1

Zelinka, M. D., Klein, S. A., & Hartmann, D. L. (2012). 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

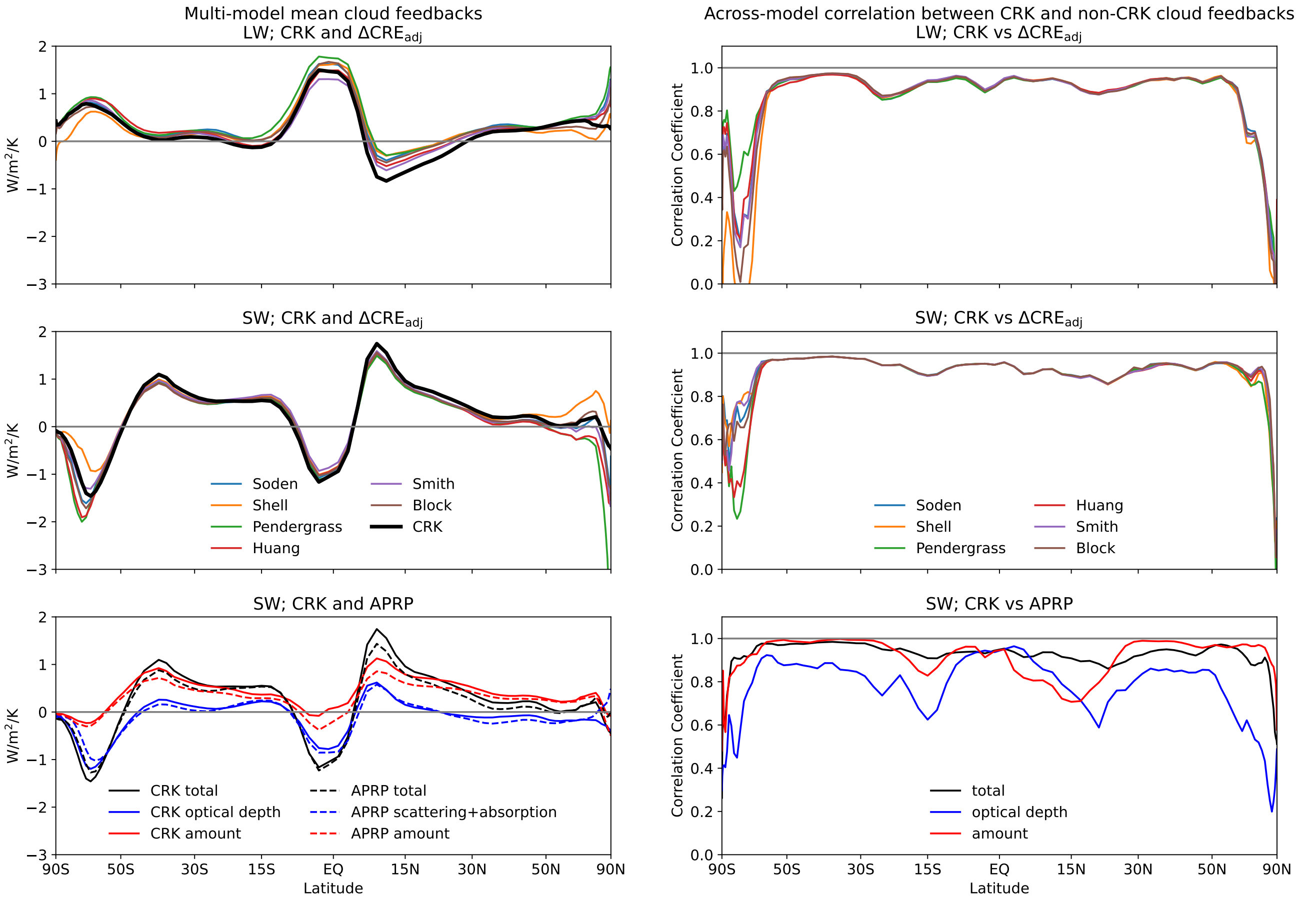

Figure S1. (left) Zonal and multi-model mean LW and SW cloud feedbacks estimated using three methodologies: cloud radiative kernels (CRK; Zelinka et al., 2012), adjusted change in cloud radiative effect (ΔCRE_{adj}; Soden et al., 2008; Shell et al., 2008), and approximate partial radiative perturbation (APRP; Taylor et al., 2007). Six estimates of ΔCRE_{adj} are shown, each using a different radiative kernel identified in the caption on row 2. (right) Across-model correlation between CRK-derived and non-CRK-derived zonal mean cloud feedbacks. The CRK-derived SW cloud feedback is further broken down into optical depth and amount components, which are compared to the APRP-derived SW scattering plus absorption component and amount component, respectively.

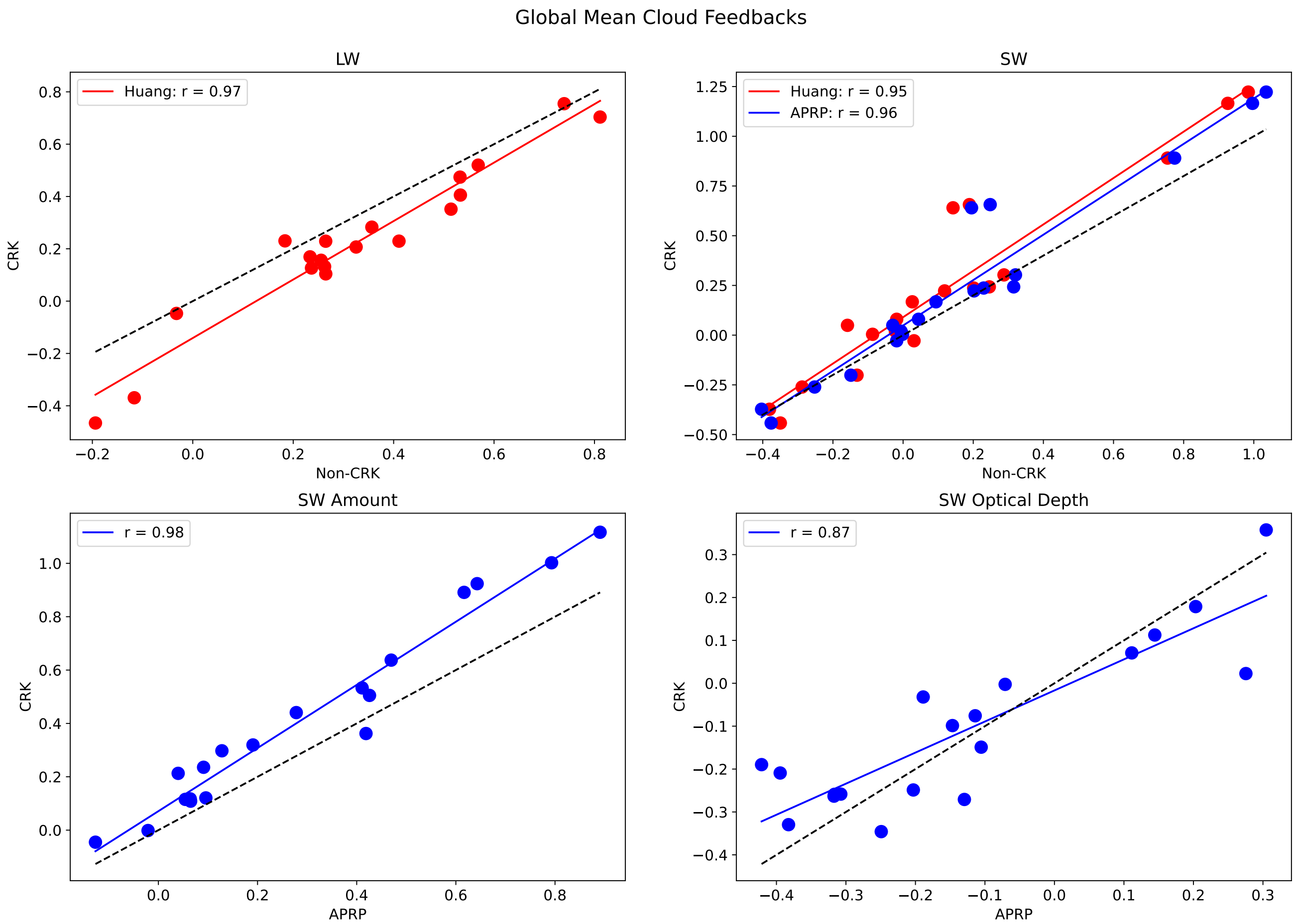

Figure S2. Global mean LW and SW cloud feedbacks estimated using the CRK method scattered against those estimated using non-CRK techniques. For clarity, we show only one of the six estimates of ΔCRE_{adj}, that derived using the kernels of Huang et al. (2017).

Cloud Feedback Components		Amount		Altitude		Optical Depth	
		Ocean	Land	Ocean	Land	Ocean	Land
90˚S-90˚N	High	N/A	4	1	1	N/A	N/A
	Low	N/A	4	1	1	N/A	N/A
30˚S-30˚N	High	Asc 3 / Dsc 2	N/A	N/A	N/A	Asc 3 / Dsc 2	4
	Low	Asc 3 / Dsc 2	N/A	N/A	N/A	Asc 3 / Dsc 2	4
30˚-40˚N/S	High	8	N/A	N/A	N/A	7	7
	Low	5	N/A	N/A	N/A	6	6
40˚-60˚N/S	High	8	N/A	N/A	N/A	7	7
	Low	5	N/A	N/A	N/A	6	6
60˚-70˚N/S	High	8	N/A	N/A	N/A	7	7
	Low	5	N/A	N/A	N/A	6	6
70˚-90˚N/S	High	8	N/A	N/A	N/A	7	7
	Low	5	N/A	N/A	N/A	6	6

Assessed
1. Global High ALT
2. Tropical Ocean Descent Low AMT + TAU
3. Anvil
4. Global Land AMT
5. Middle latitude Low AMT
6. Extratropical Low TAU

Unassessed
1. Global Low ALT
2. Tropical Ocean Descent High AMT+TAU
3. Tropical Ocean Ascent Low AMT+TAU
4. Tropical Land High+Low TAU
5. 60-90 Ocean Low AMT
6. 30-40/70-90 Ocean+Land Low TAU
7. 30-90 Ocean+Land High TAU
8. 30-90 Ocean High AMT
9. Global Obscuration Covariance*
10. Global Zelinka et al. (2013) Residual*

*not shown in matrix for brevity

N/A = Not Applicable

Figure S3. Matrix of assessed and unassessed cloud feedbacks. The sum of all assessed and unassessed components equals the total cloud feedback.

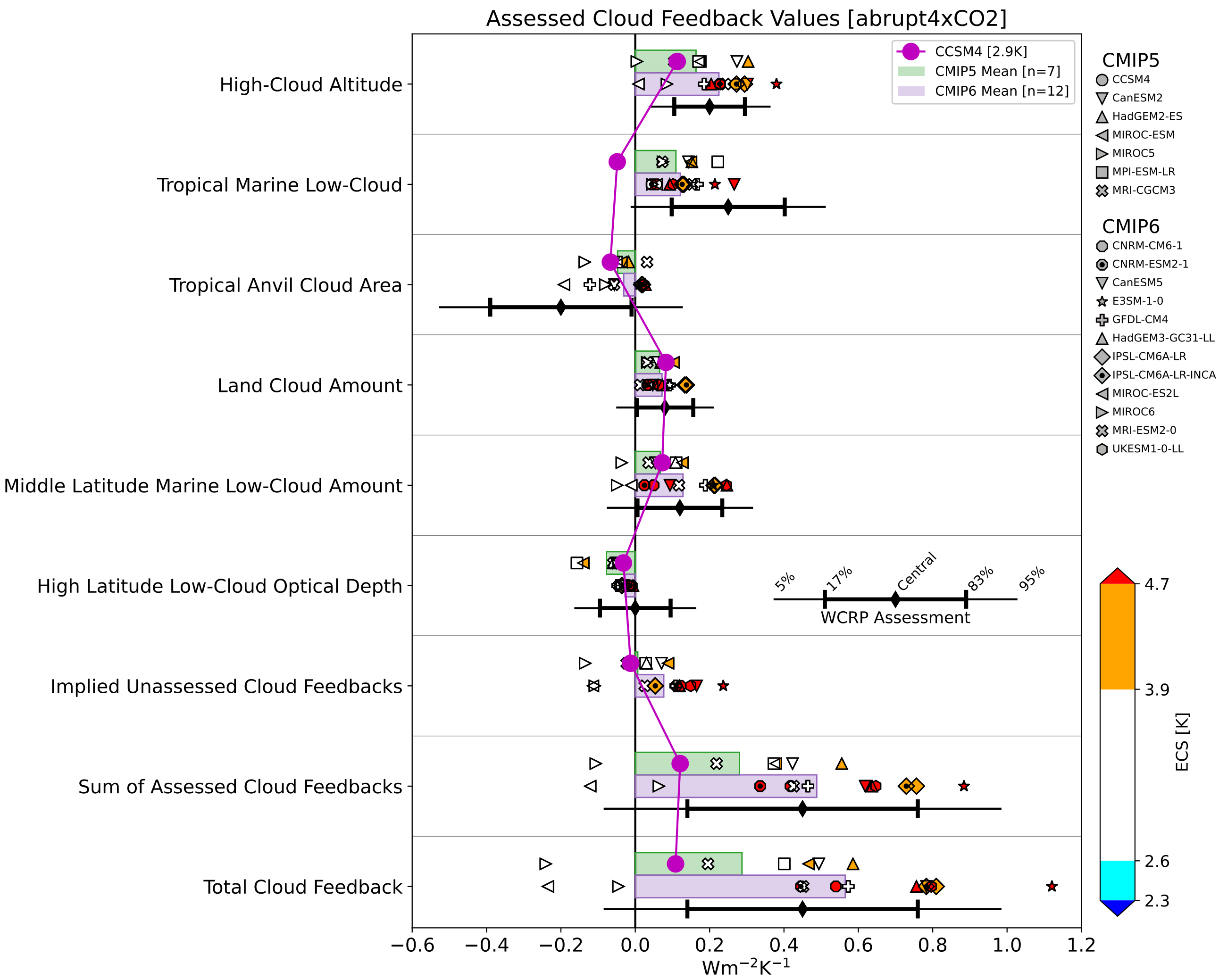

Figure S4. As in Figure 1, but highlighting CCSM4.

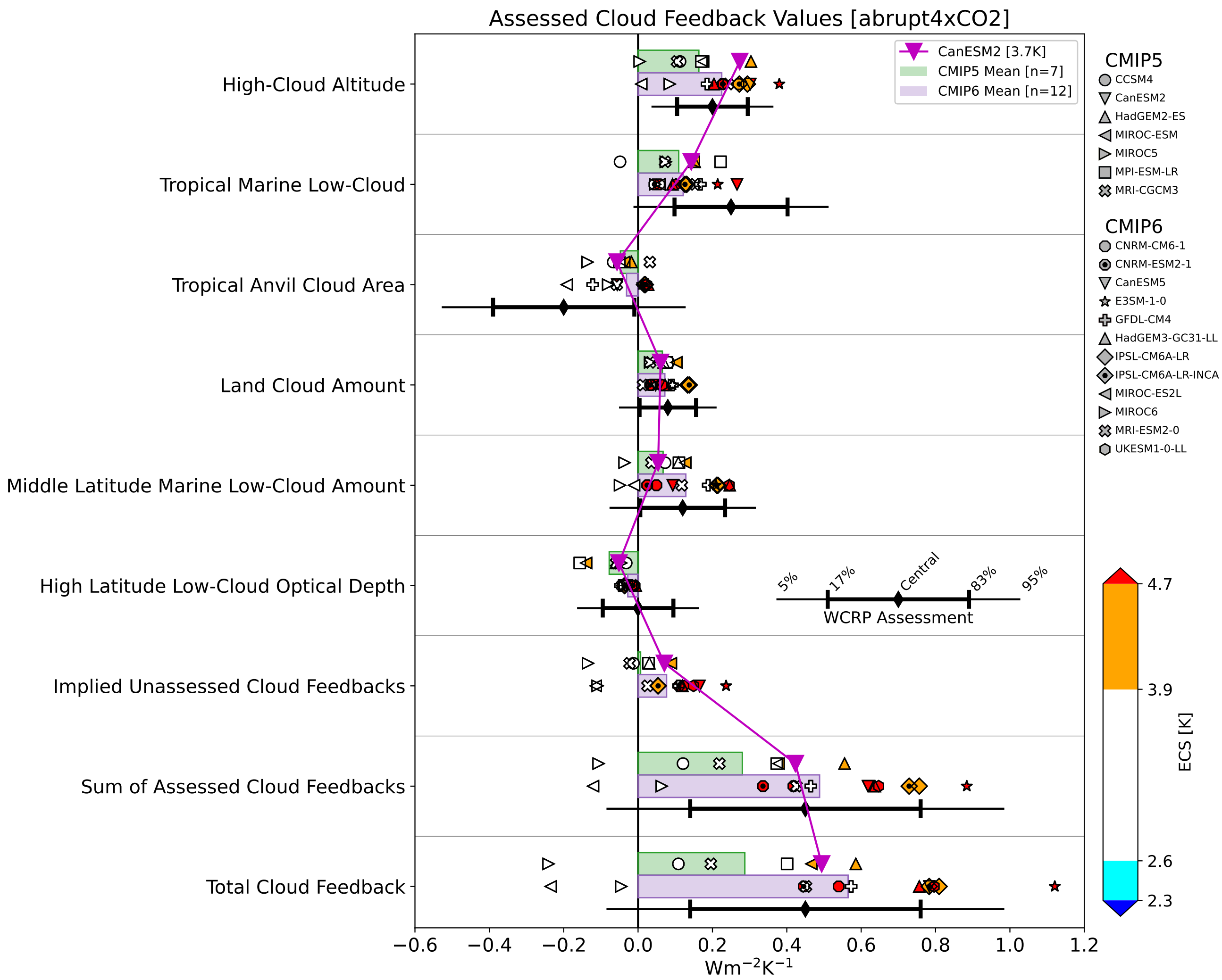

Figure S5. As in Figure 1, but highlighting CanESM2.

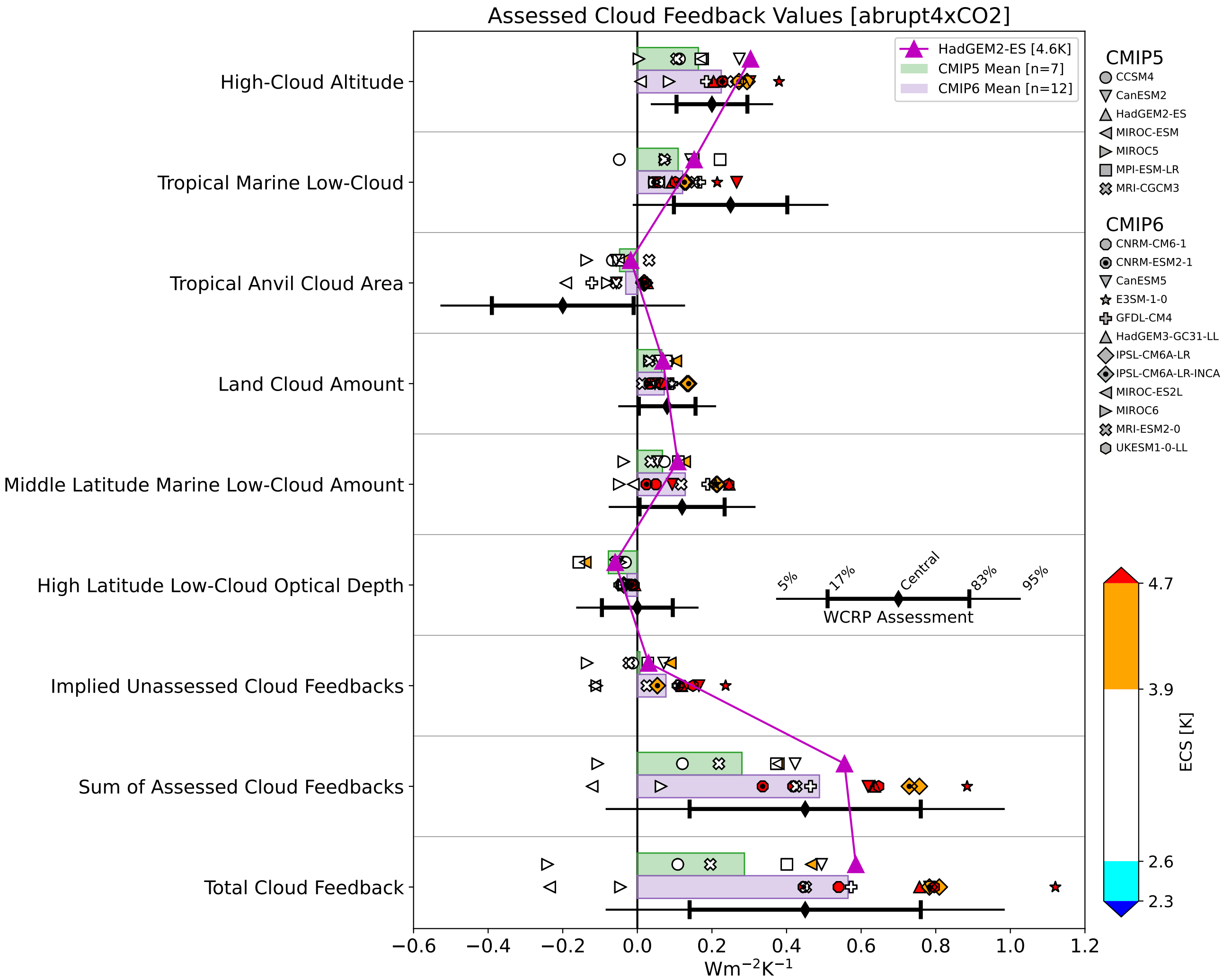

Figure S6. As in Figure 1, but highlighting HadGEM2-ES.

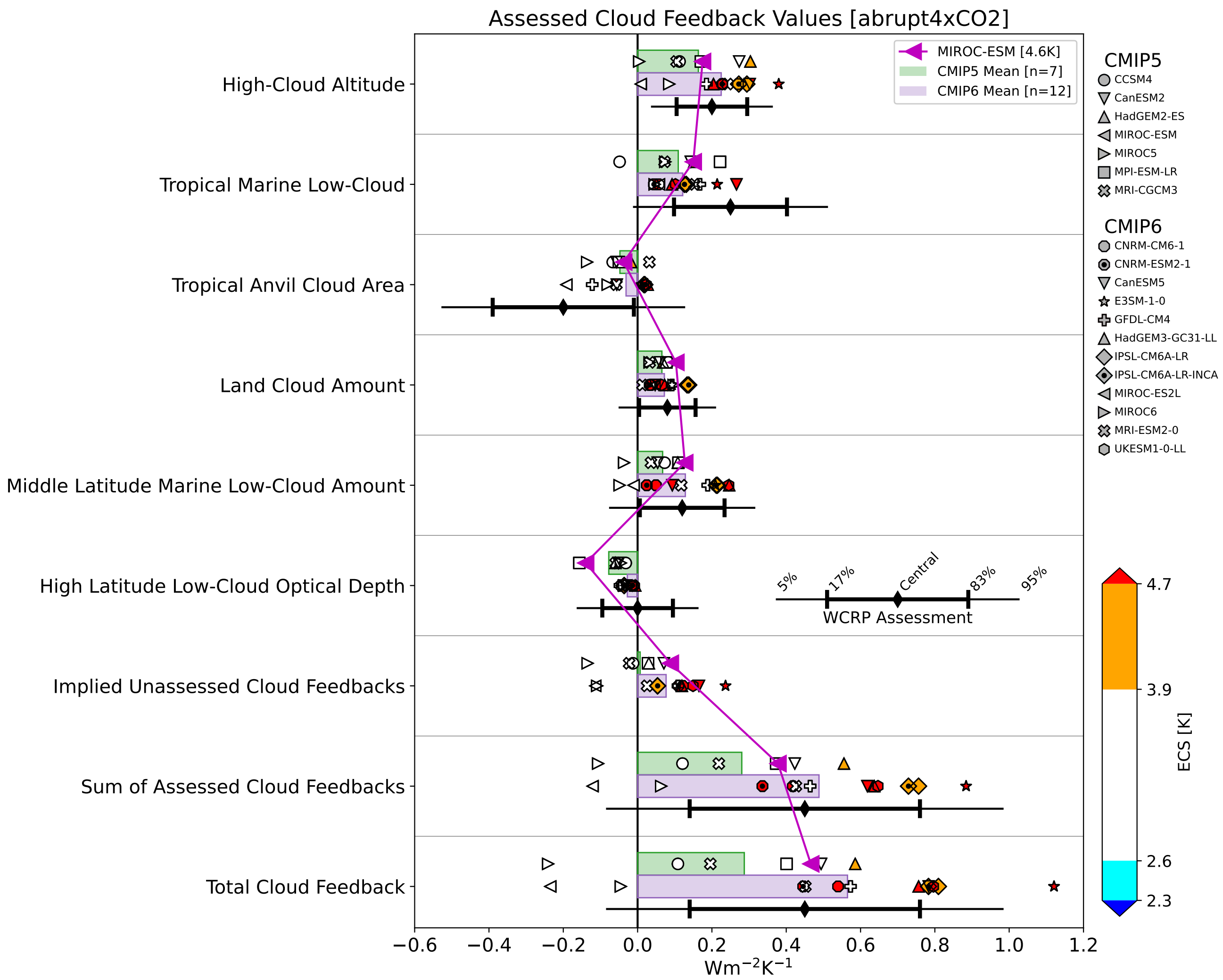

Figure S7. As in Figure 1, but highlighting MIROC-ESM.

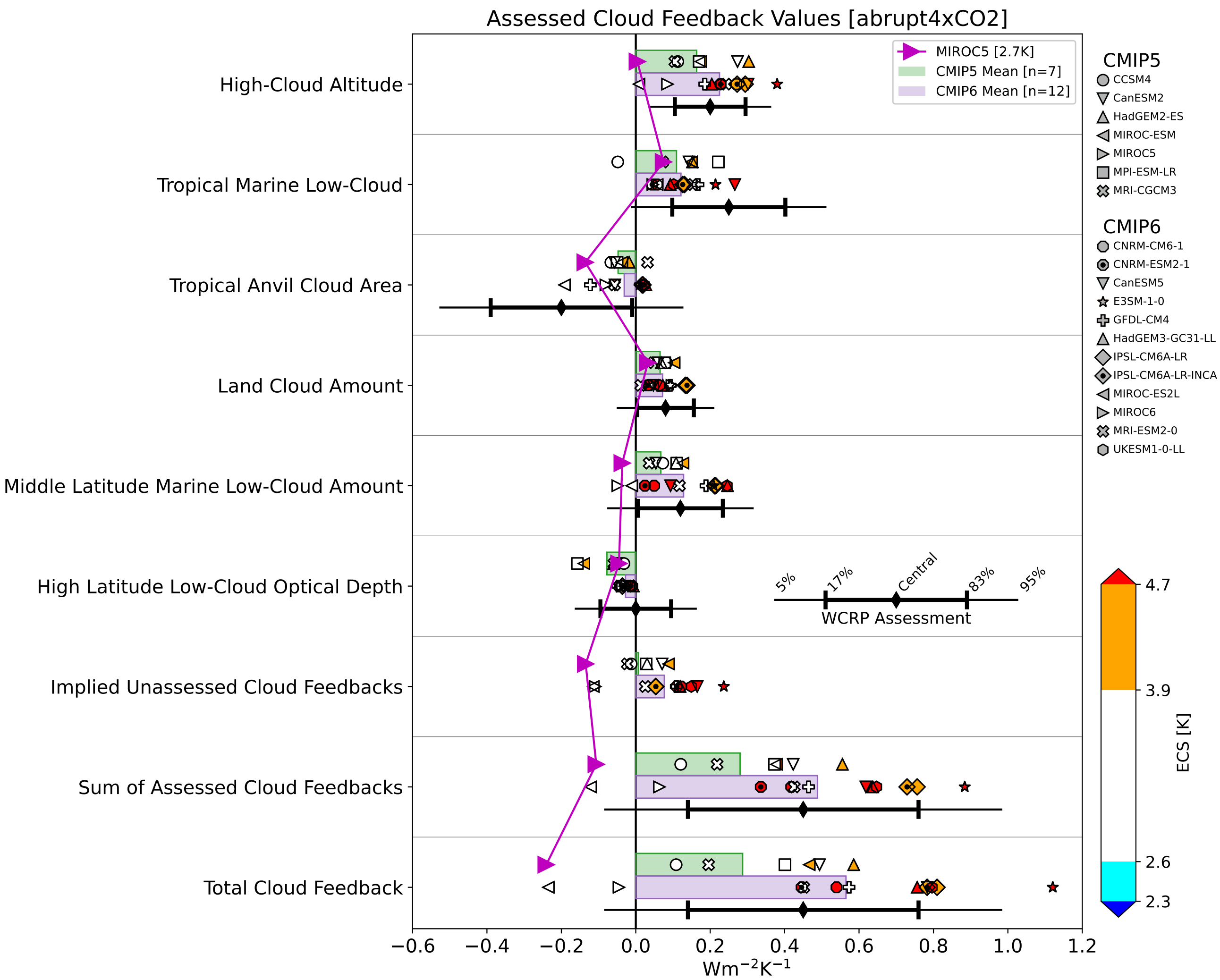

Figure S8. As in Figure 1, but highlighting MIROC5.

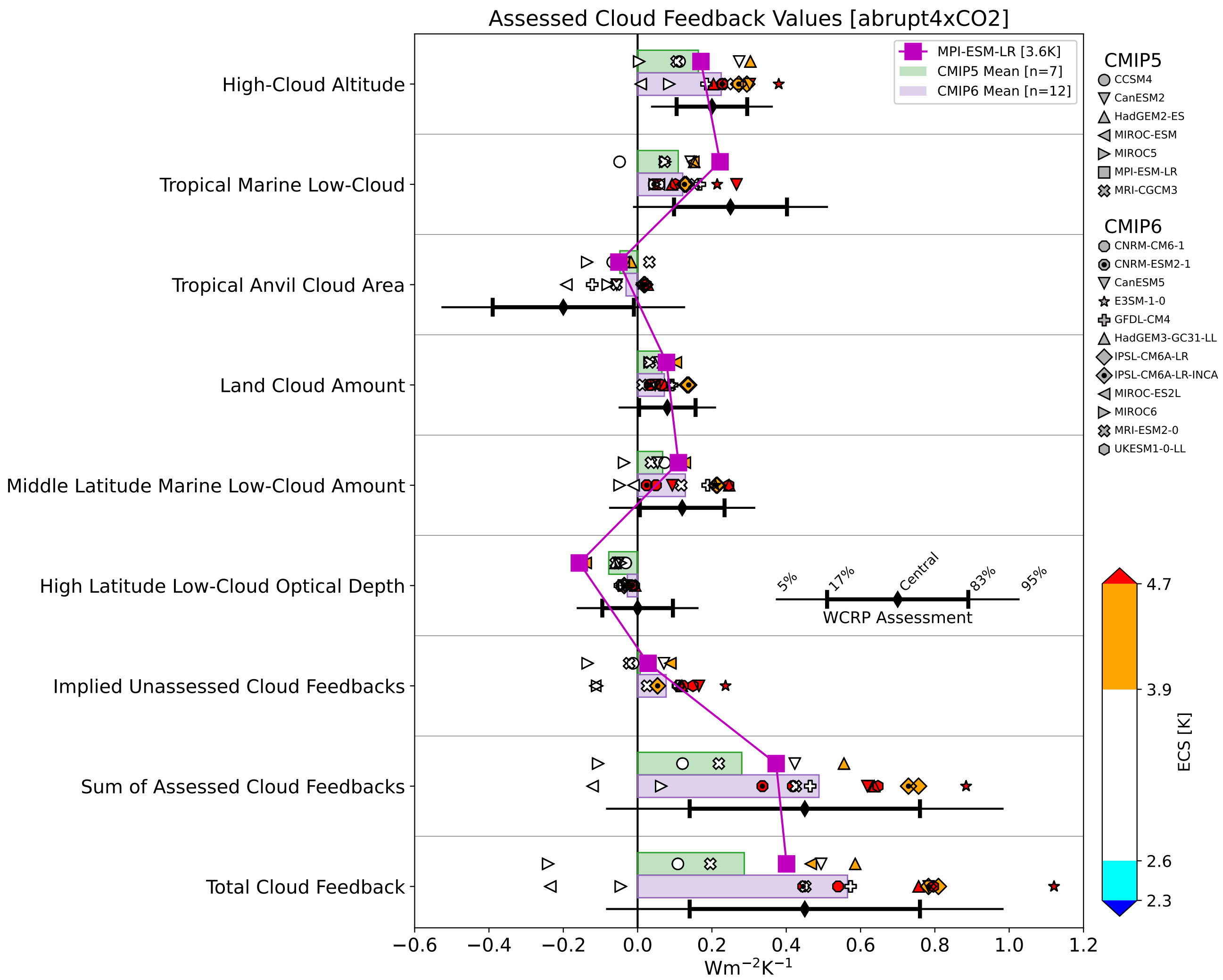

Figure S9. As in Figure 1, but highlighting MPI-ESM-LR.

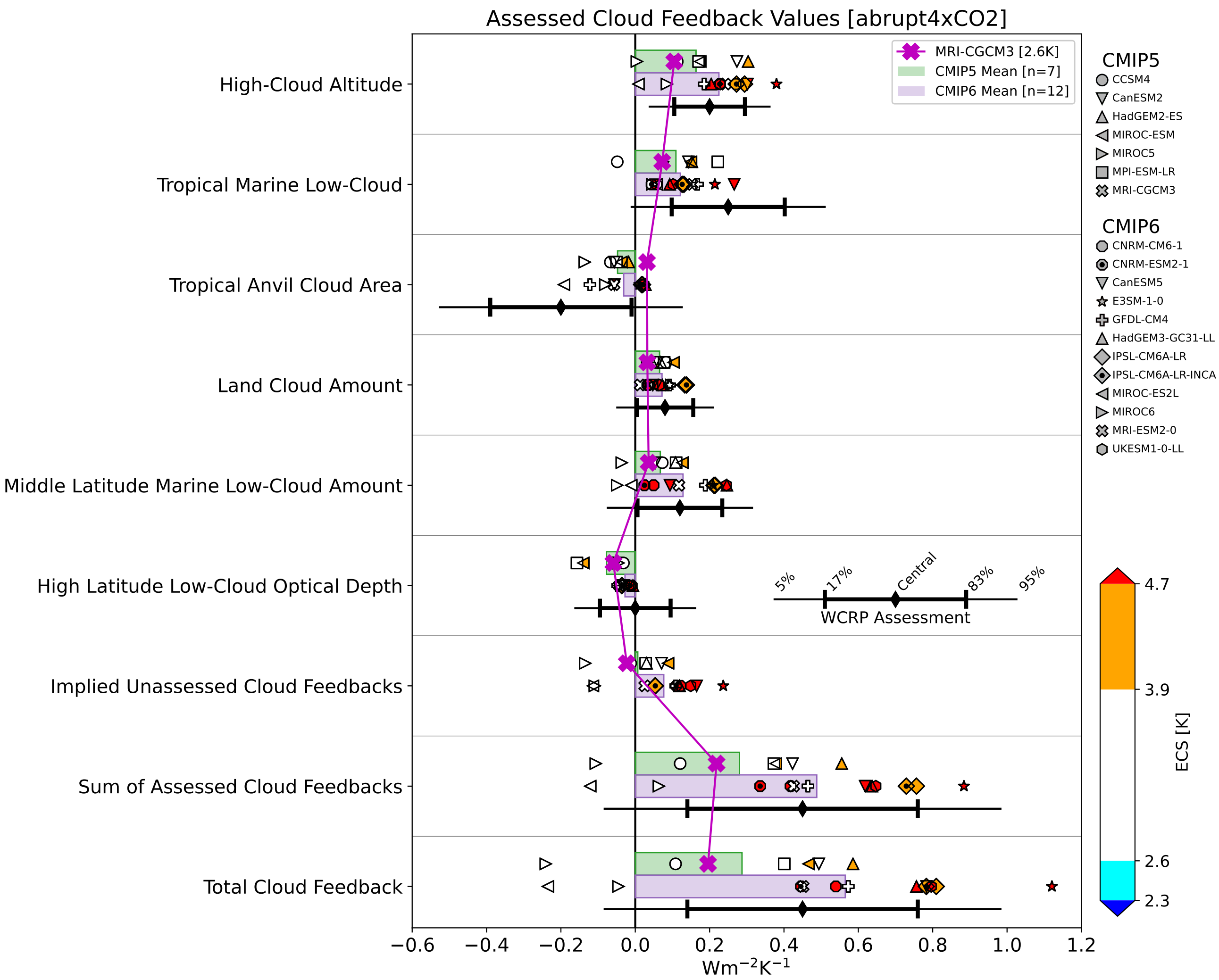

Figure S10. As in Figure 1, but highlighting MRI-CGCM3.

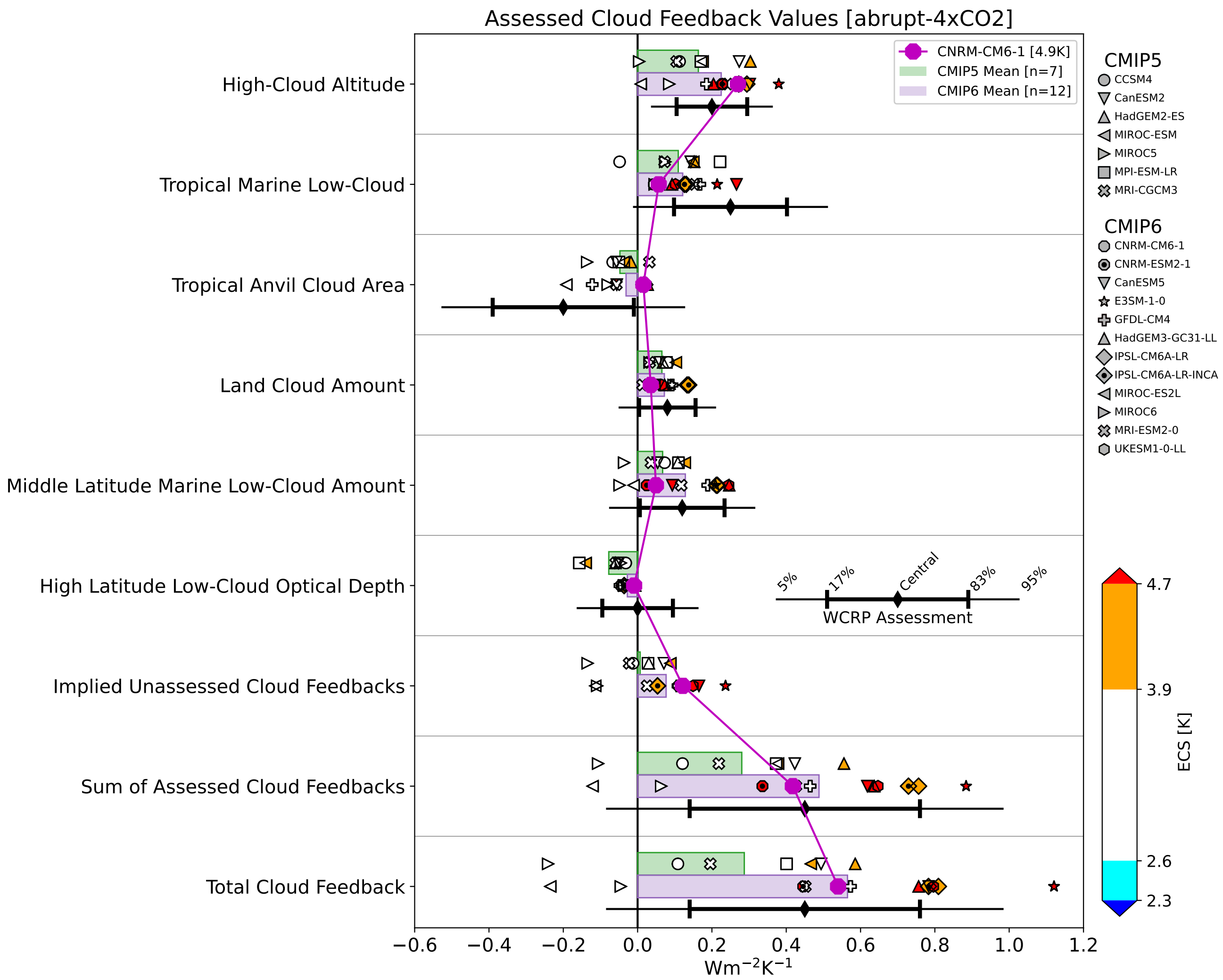

Figure S11. As in Figure 1, but highlighting CNRM-CM6-1.

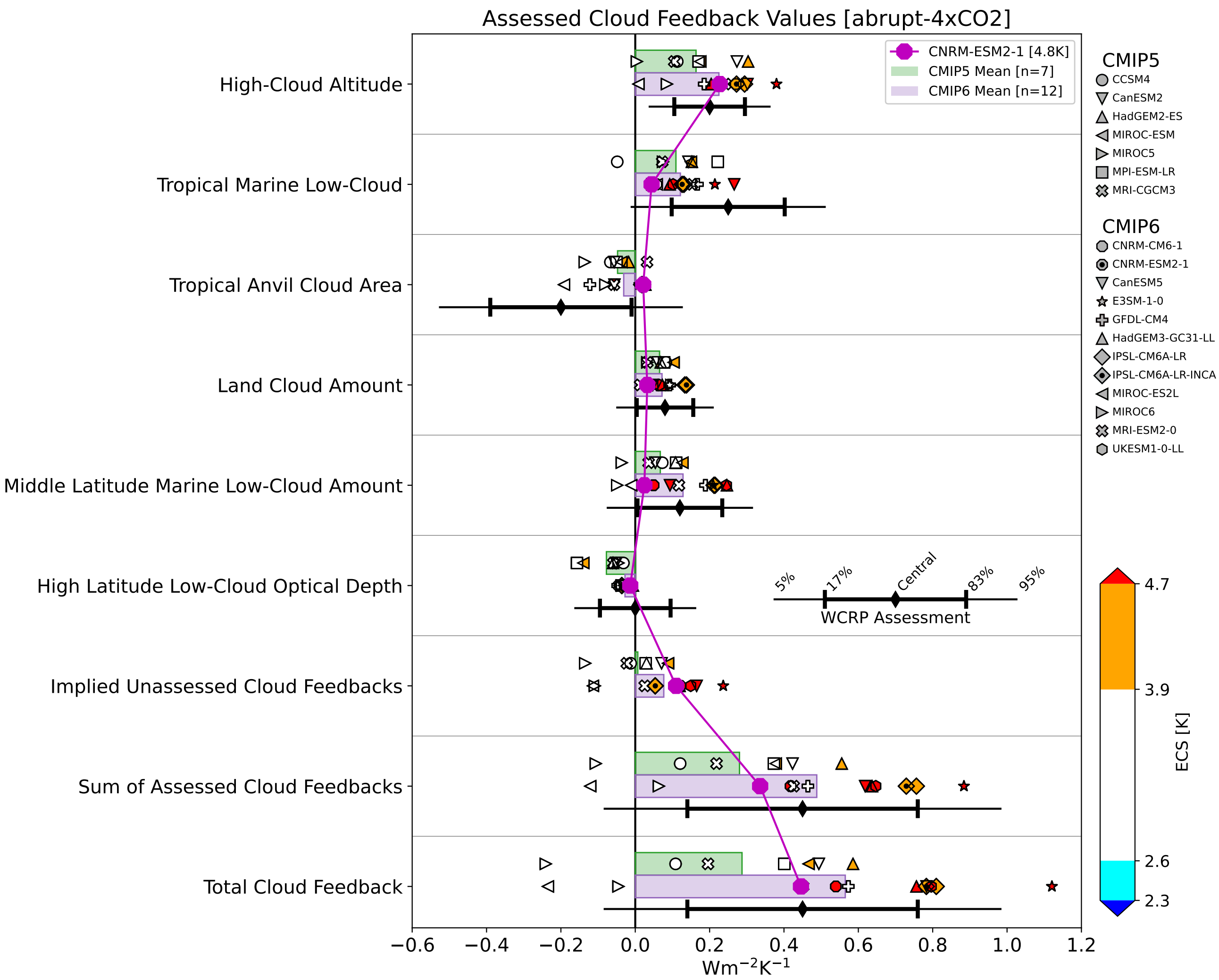

Figure S12. As in Figure 1, but highlighting CNRM-ESM2-1.

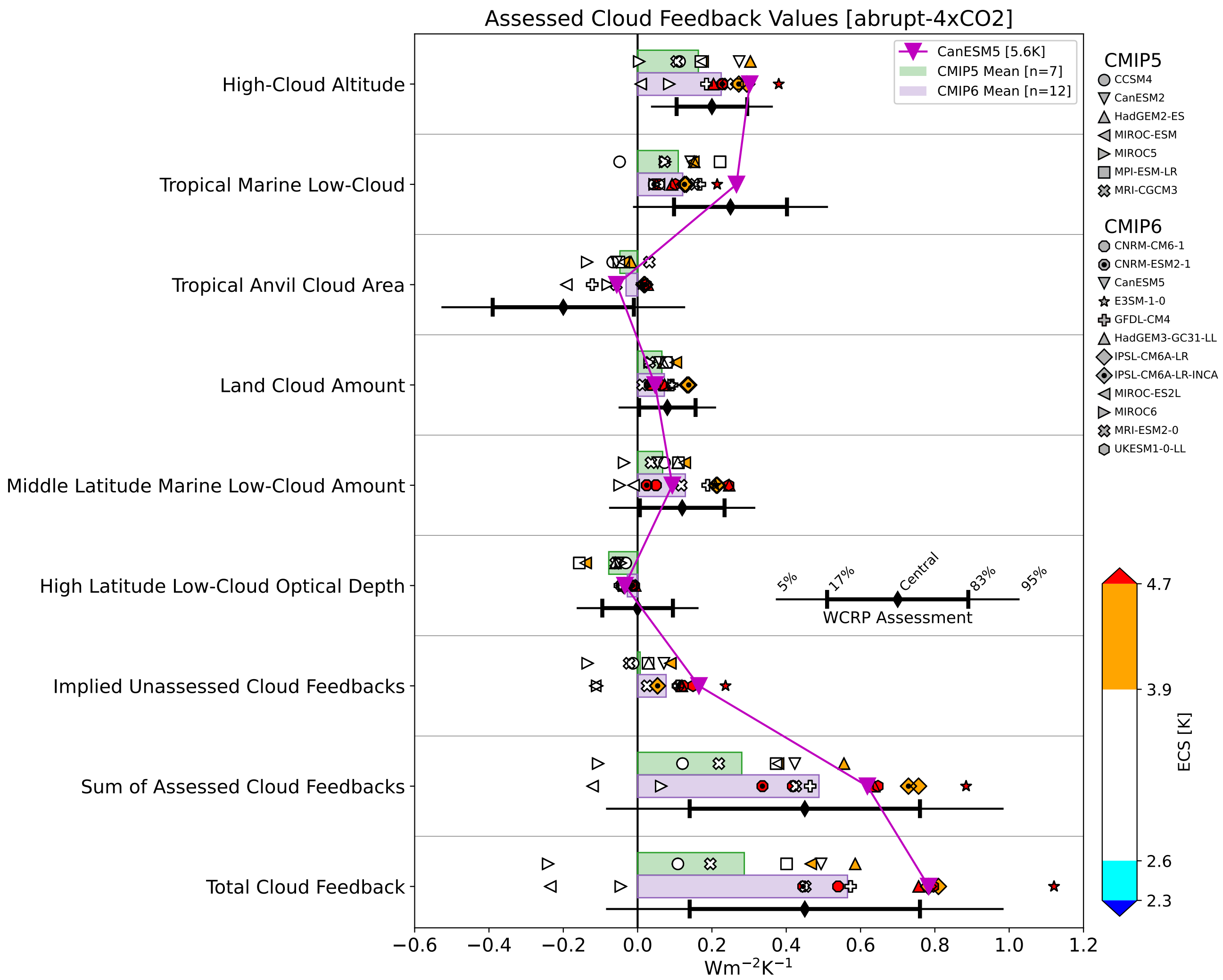

Figure S13. As in Figure 1, but highlighting CanESM5.

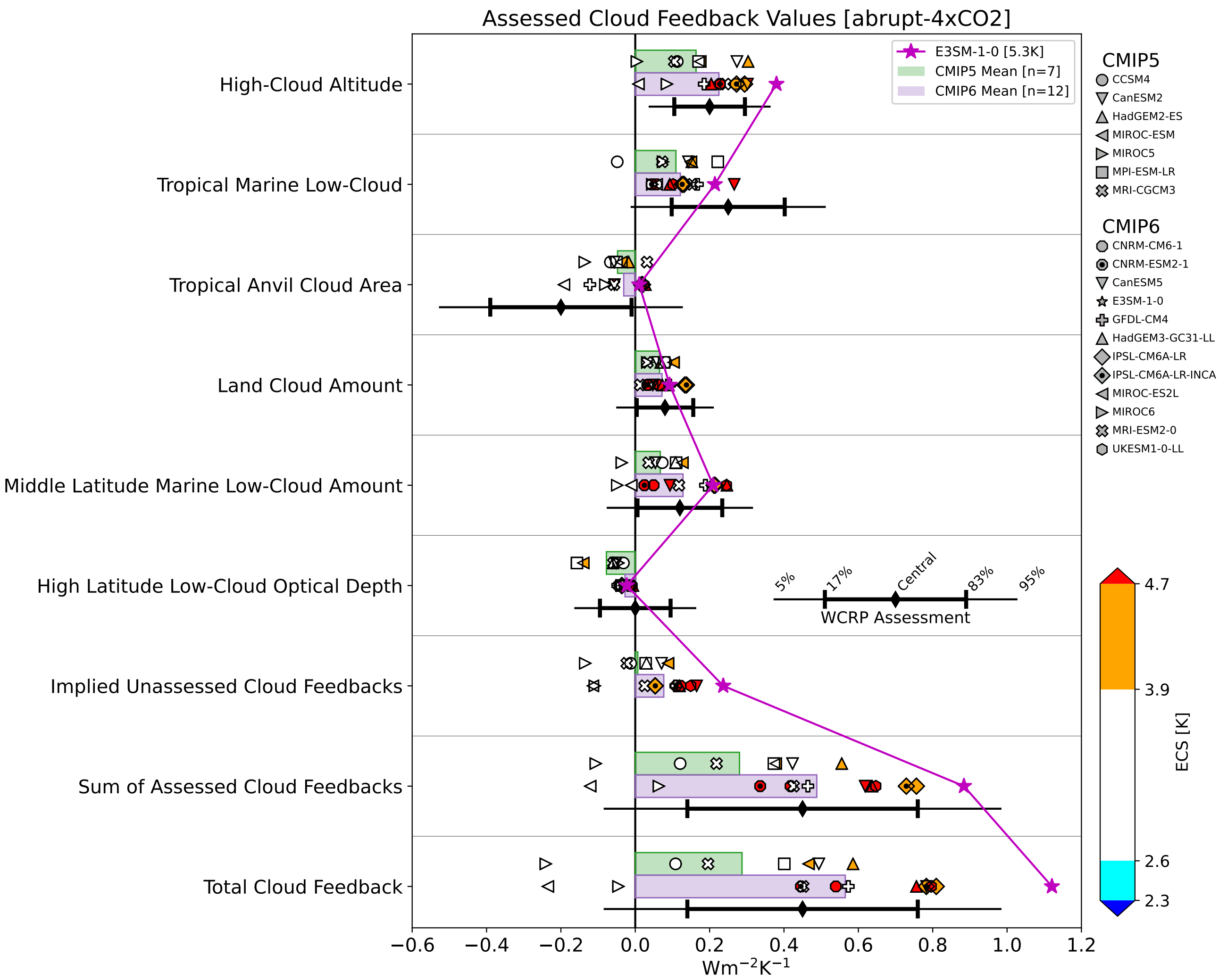

Figure S14. As in Figure 1, but highlighting E3SM-1-0.

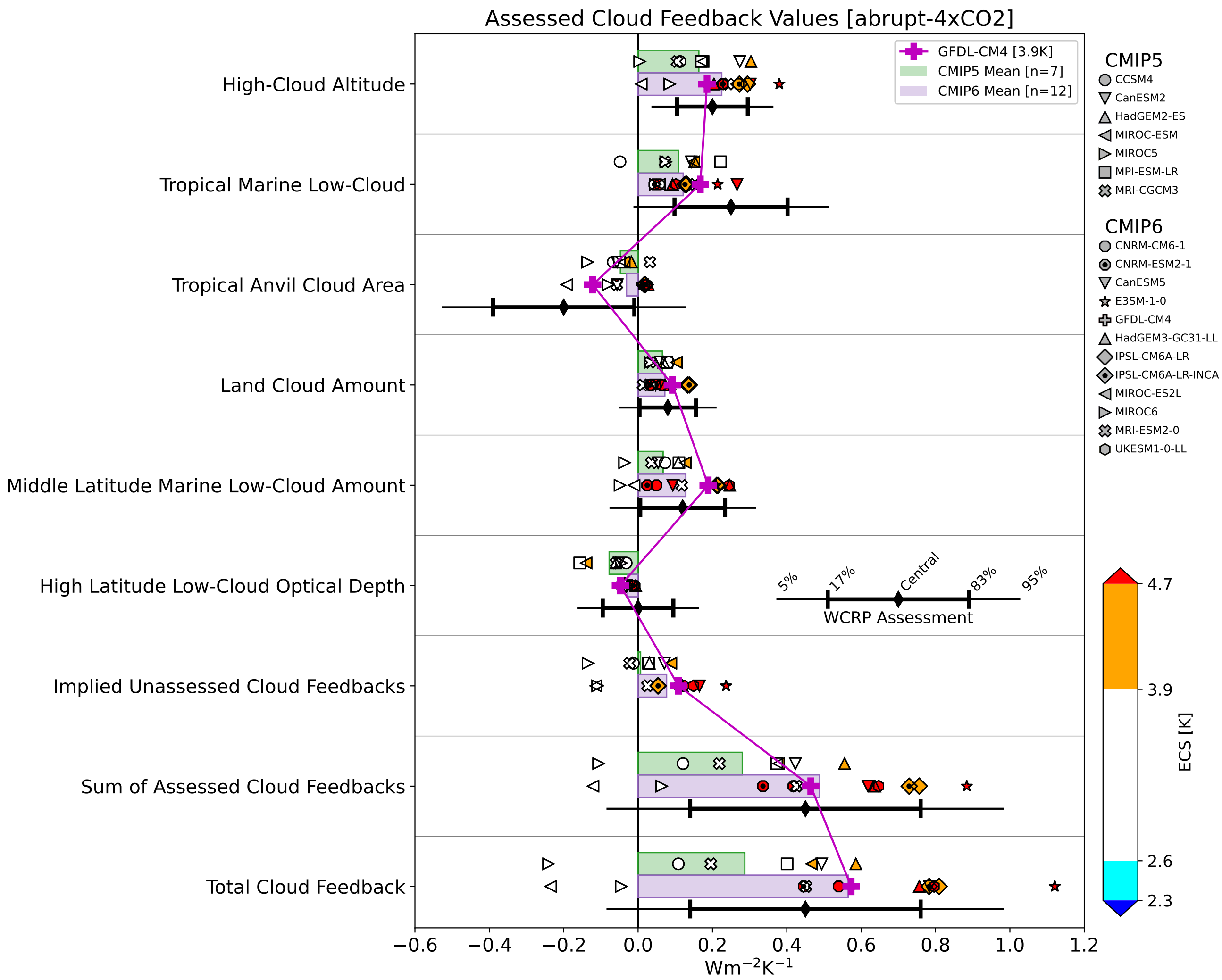

Figure S15. As in Figure 1, but highlighting GFDL-CM4.

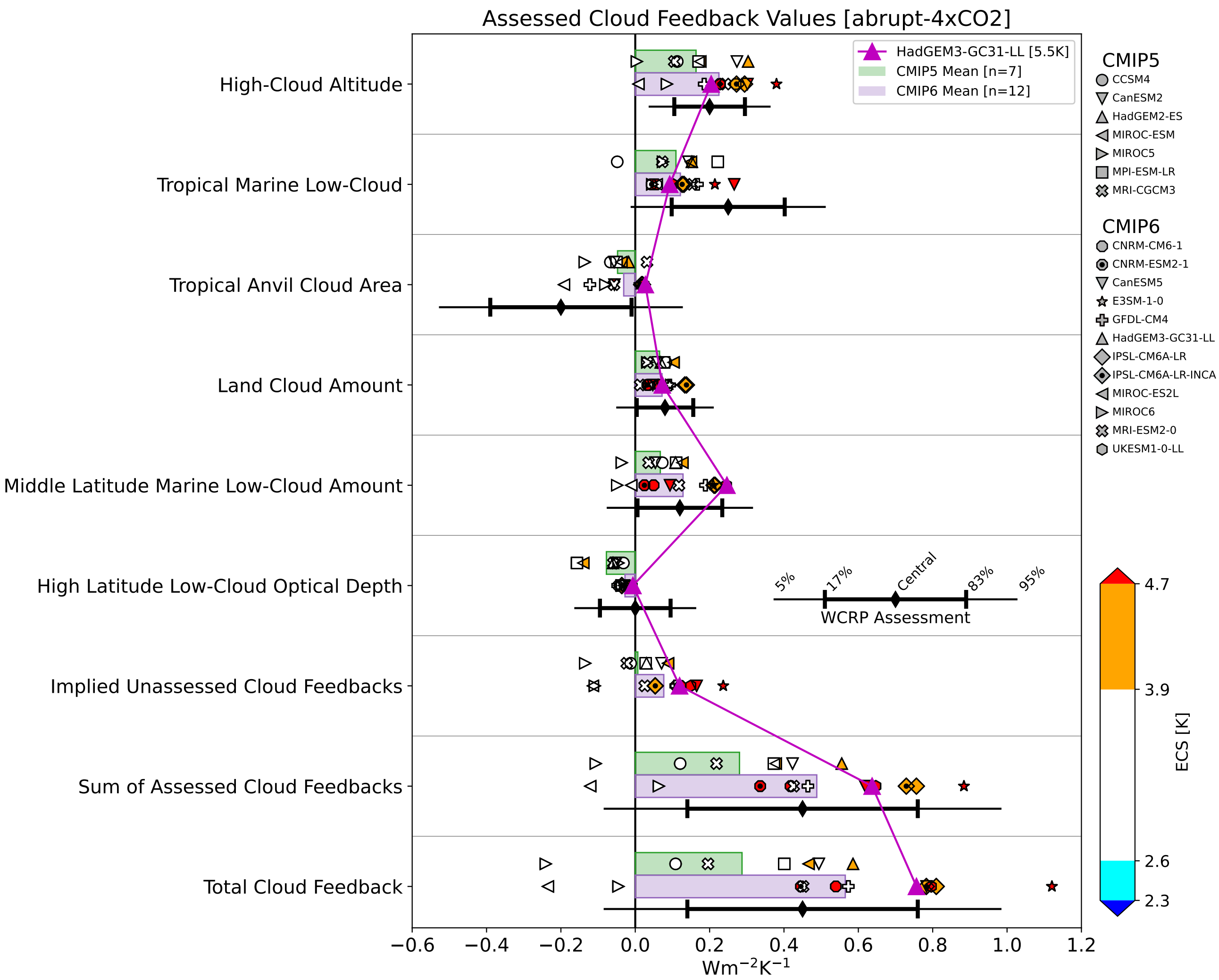

Figure S16. As in Figure 1, but highlighting HadGEM3-GC31-LL.

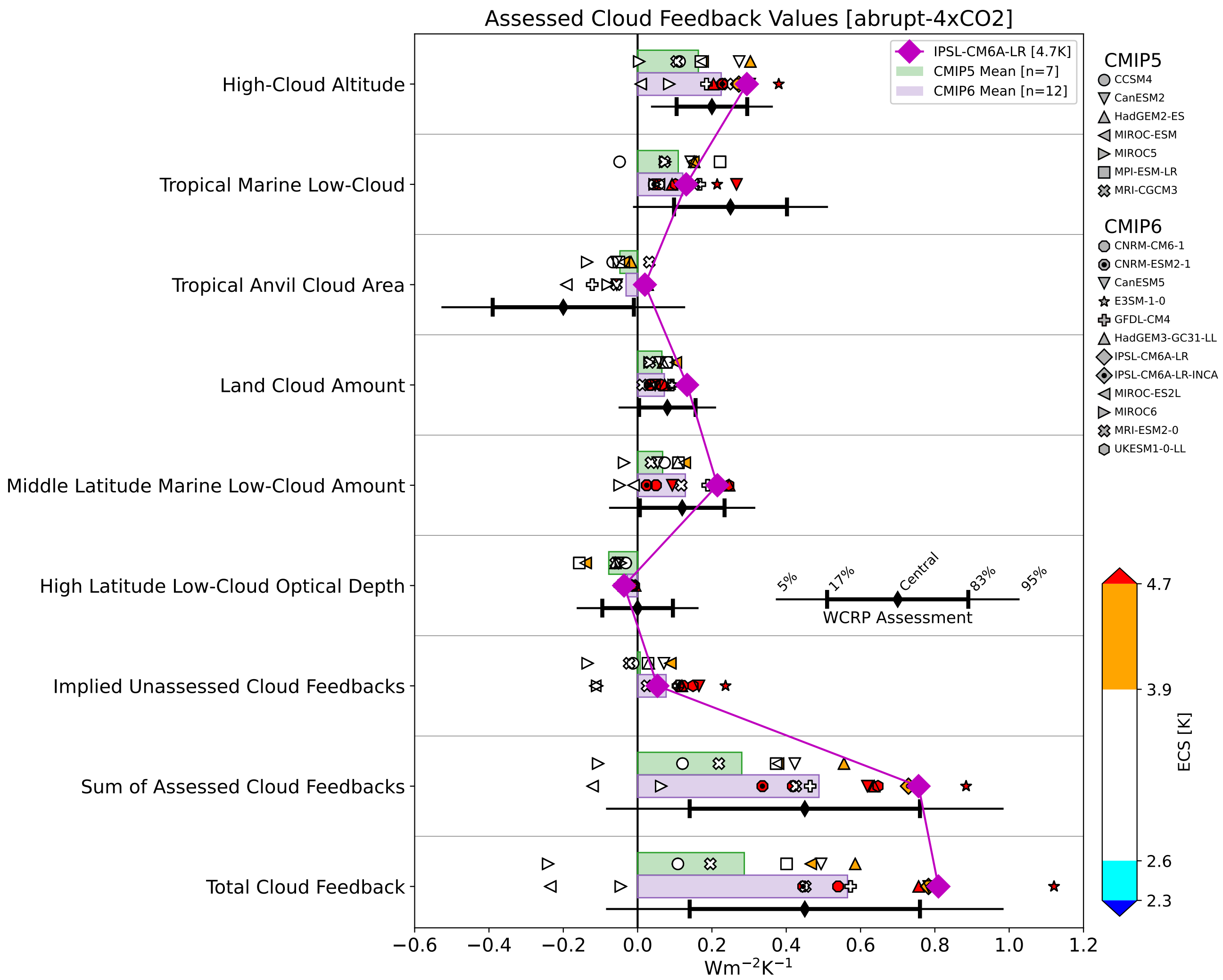

Figure S17. As in Figure 1, but highlighting IPSL-CM6A-LR.

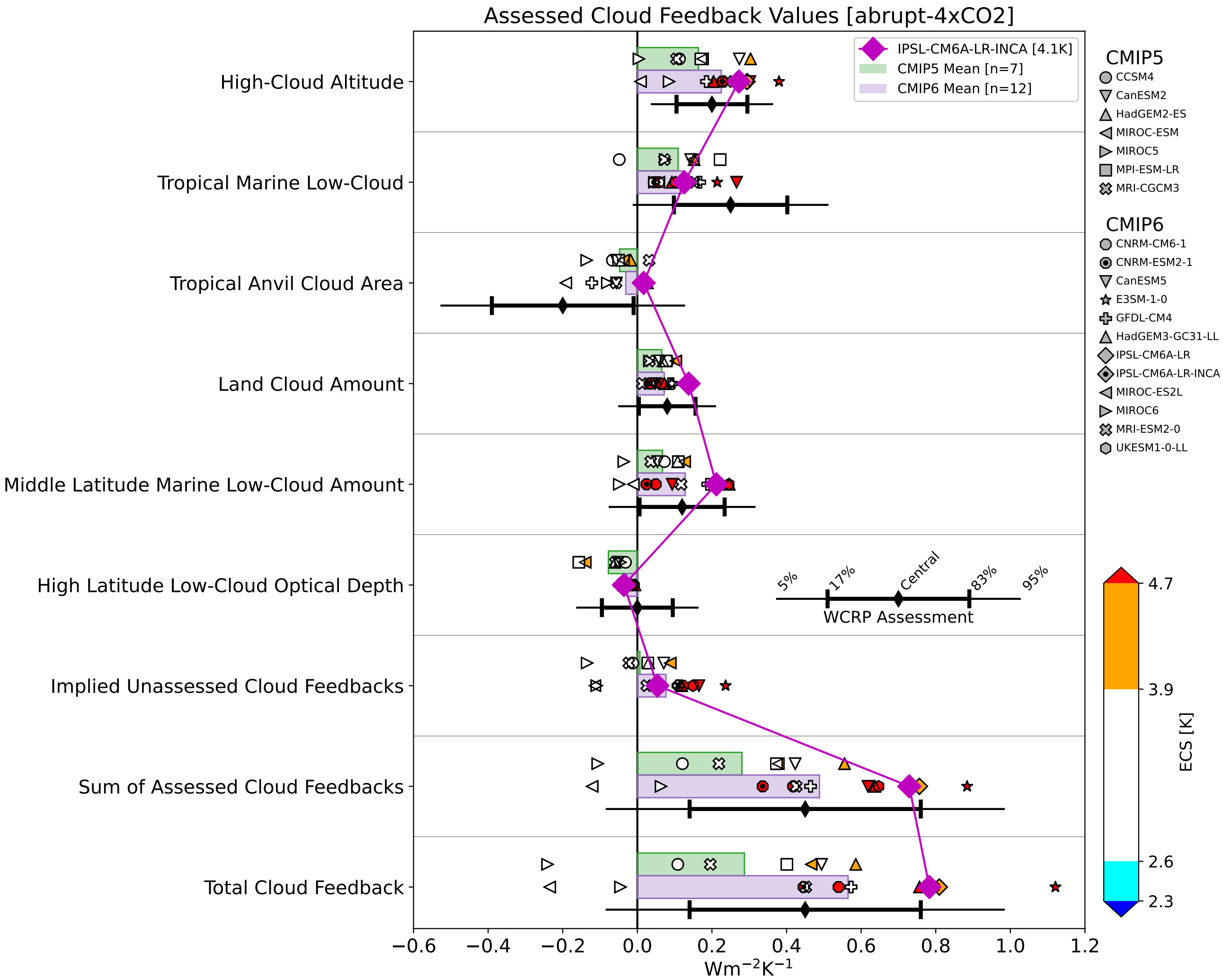

Figure S18. As in Figure 1, but highlighting IPSL-CM6A-LR-INCA.

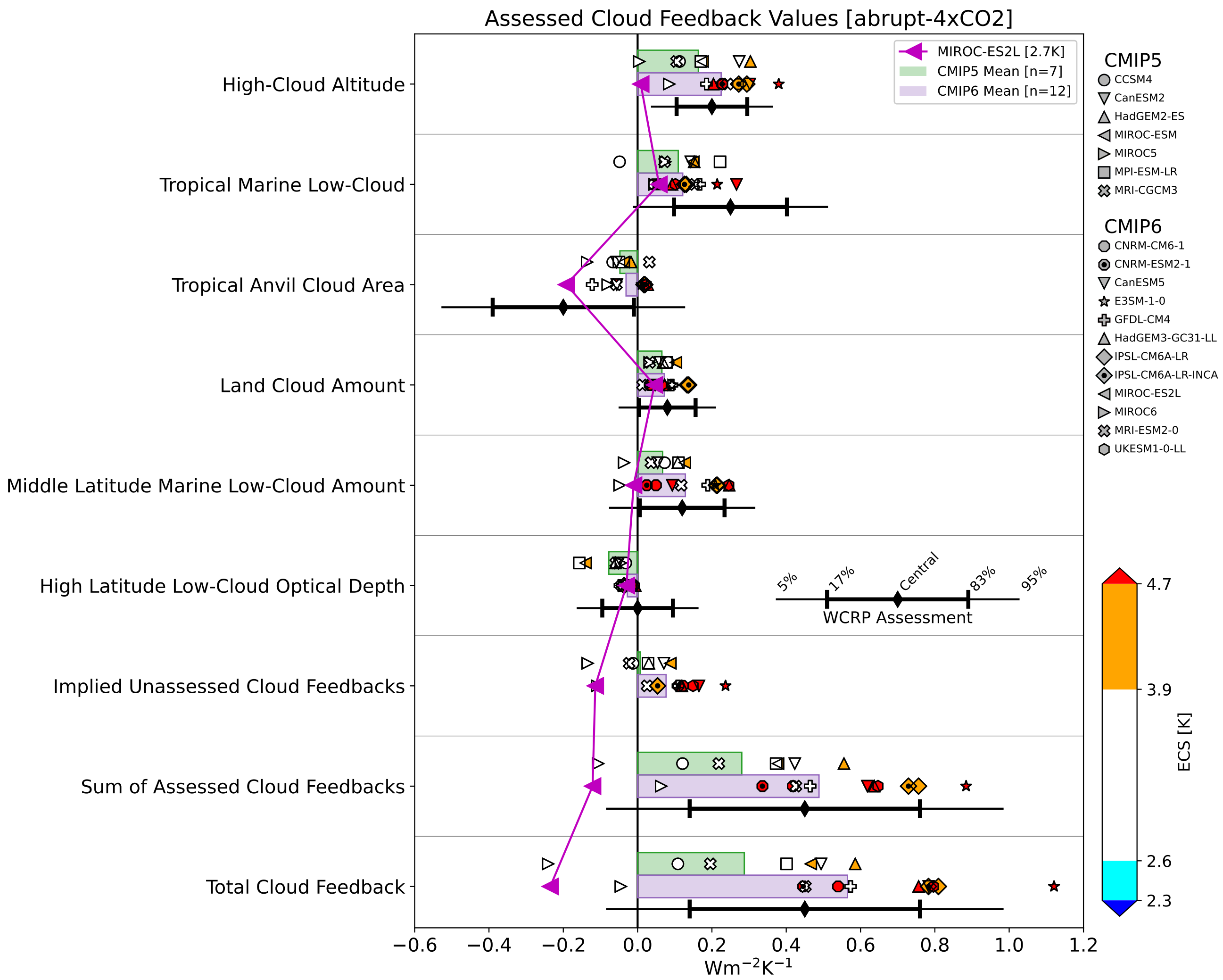

Figure S19. As in Figure 1, but highlighting MIROC-ES2L.

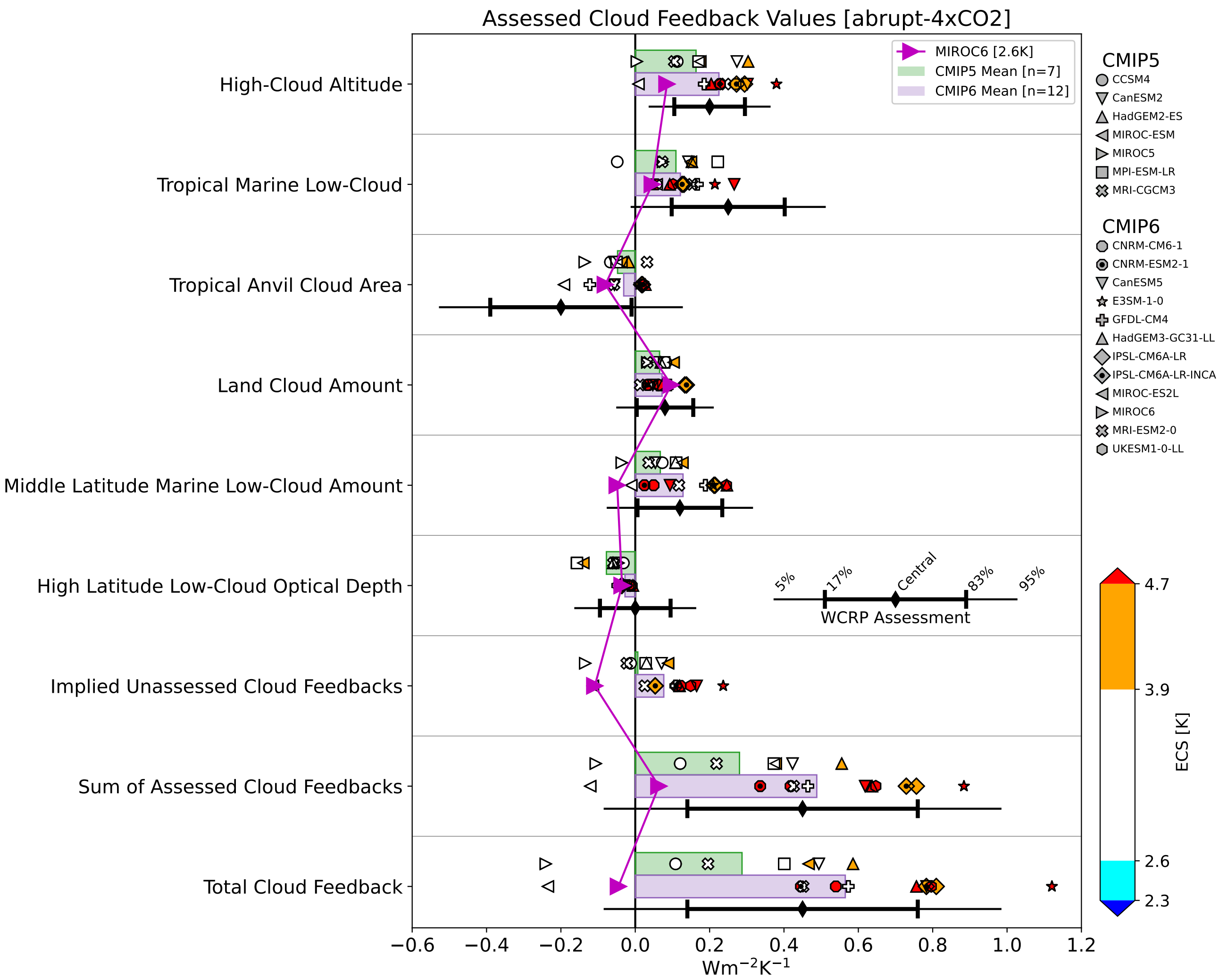

Figure S20. As in Figure 1, but highlighting MIROC6.

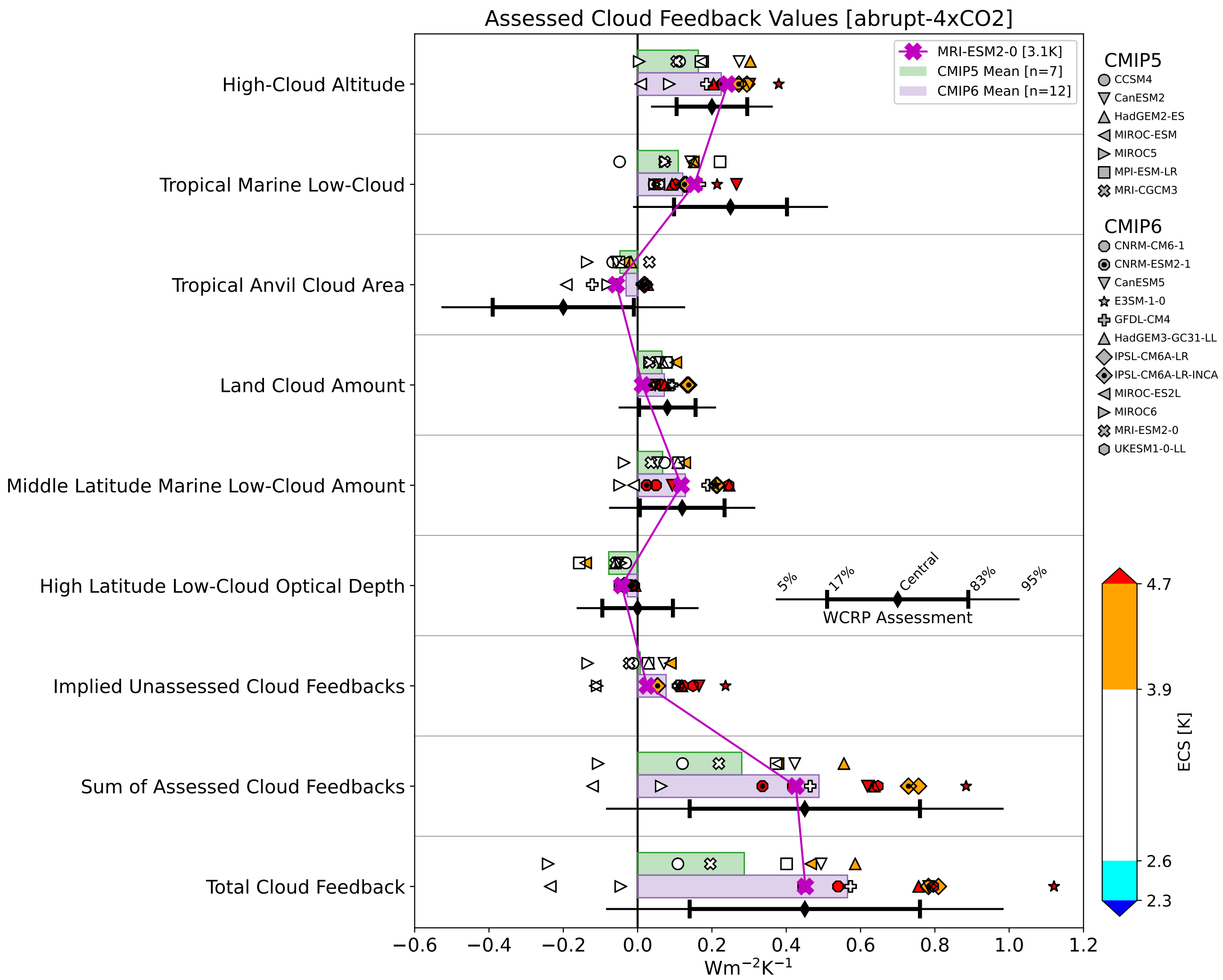

Figure S21. As in Figure 1, but highlighting MRI-ESM2-0.

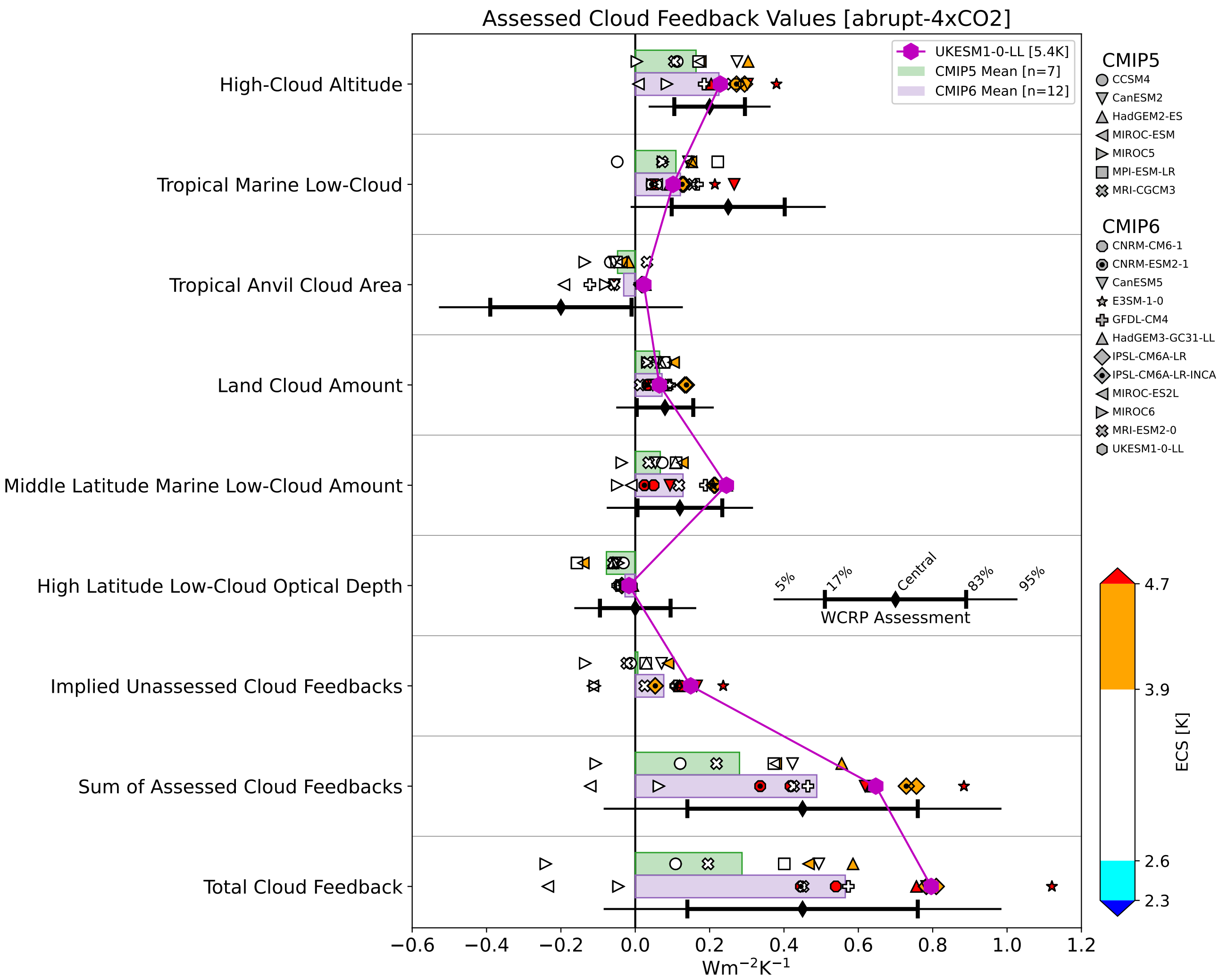

Figure S22. As in Figure 1, but highlighting UKESM1-0-LL.