Correspondence to:

B. Medeiros,
brianpm@ucar.edu

Author Contributions:

Conceptualization: Brian Medeiros, Jonah Shaw, Jennifer E. Kay
Data curation: Jonah Shaw, Isaac Davis
Formal analysis: Brian Medeiros
Visualization: Brian Medeiros
Writing – original draft: Brian Medeiros
Writing – review & editing: Brian Medeiros, Jonah Shaw, Jennifer E. Kay

Assessing Clouds Using Satellite Observations Through Three Generations of Global Atmosphere Models

Brian Medeiros[1] , **Jonah Shaw[2]** , **Jennifer E. Kay[2]** , **and Isaac Davis[1]**

[1]National Center for Atmospheric Research, Boulder, CO, USA, [2]Department of Atmospheric and Oceanic Sciences, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Boulder, CO, USA

Abstract Clouds are parameterized in climate models using quantities on the model grid-scale to approximate the cloud cover and impact on radiation. Because of the complexity of processes involved with clouds, these parameterizations are one of the key challenges in climate modeling. Differences in parameterizations of clouds are among the main contributors to the spread in climate sensitivity across models. In this work, the clouds in three generations of an atmosphere model lineage are evaluated against satellite observations. Satellite simulators are used within the model to provide an appropriate comparison with individual satellite products. In some respects, especially the top-of-atmosphere cloud radiative effect, the models show generational improvements. The most recent generation, represented by two distinct branches of development, exhibits some regional regressions in the cloud representation; in particular the southern ocean shows a positive bias in cloud cover. The two branches of model development show how choices during model development, both structural and parametric, lead to different cloud climatologies. Several evaluation strategies are used to quantify the spatial errors in terms of the large-scale circulation and the cloud structure. The Earth mover's distance is proposed as a useful error metric for the passive satellite data products that provide cloud-top pressure-optical depth histograms. The cloud errors identified here may contribute to the high climate sensitivity in the Community Earth System Model, version 2 and in the Energy Exascale Earth System Model, version 1.

1. Introduction

Assessments of climate sensitivity, the warming due to a doubling of CO_2, in the sixth phase of CMIP (see Table A1 for all abbreviations) indicate that some climate models warm more than earlier models (e.g., Meehl et al., 2020; Zelinka et al., 2020). Some argue that a group of especially "hot" models are outside the plausible range of warming (Hausfather et al., 2022; Tokarska et al., 2020). Recent work indicates that a third of CMIP6 models may, in fact, have sensitivity of greater than 5K (He et al., 2022). Several of the high-sensitivity models are versions of, or closely related to, the Community Earth System Model (CESM).

Climate sensitivity in CESM has increased through the last three generations as the atmospheric component, CAM, has progressed from CAM4 with a sensitivity of 3.0 K to CAM5 at 4.1 K and the most recent version, CAM6, at 5.2 K. From all indications, this increase arises mainly from cloud feedbacks becoming more positive through the generations. This is indicated in reports comparing CAM4 and CAM5 by Gettelman et al. (2012) while Bacmeister et al. (2020) and Gettelman et al. (2019) discuss cloud feedbacks and climate sensitivity in CAM6. Bacmeister et al. (2020) shows that in very long simulations, the climate sensitivity of CESM2 is, in fact, larger than 5.2 K due to methodological reasons as well as a strong positive feedback over the Southern Ocean that emerges after substantial warming. Bjordal et al. (2020) also focus on the Southern Ocean, showing that the negative "cloud phase feedback" (clouds becoming more reflective as they become less glaciated) reduces and is overtaken by the positive cloud amount feedback once they are dominated by the liquid phase. Similarly, changes in the amount of ice in the present-day climate is reported to be a major factor in the increase in estimates of climate sensitivity from CMIP5 to CMIP6 (Zelinka et al., 2020).

Those studies focused on changes in clouds that impact climate sensitivity, but they do not provide much in terms of evaluating the simulated clouds. Evaluating climate model representations of clouds is encumbered by a number of factors that are all grounded in the reality that clouds are complicated, multi-faceted features that interact with numerous processes across many scales. Observing clouds is similarly complicated by their complex nature, and must necessarily include many subjective choices that combine with instrument and sampling

uncertainties. In spite of the challenges, the past several decades have seen unprecedented advancements in long-term satellite-based cloud observations (e.g., Stubenrauch et al., 2013). There are now several data sets that provide close to or more than two decades of high-quality cloud observations. To facilitate comparison with these observations, "satellite simulators" that emulate satellite products based on the model's atmospheric state have been developed for climate models and are provided by COSP (Bodas-Salcedo et al., 2011). The use of satellite simulators has spurred systematic evaluation of climate model clouds over the past decade. Kay et al. (2012) provided a detailed comparison of clouds in CAM4 and CAM5, and showed that CAM5 produced substantially improved clouds over the older CAM4. Klein et al. (2013) used the ISCCP simulator included in COSP to assess the improvement in clouds from CMIP3 to CMIP5, and concluded that on the whole climate model clouds showed an improvement across the generations by increasing cloud cover and reducing reflectivity to better match observations. Nam et al. (2012), however, stress that CMIP5 models still exhibit a pervasive "too-few-too-bright" bias, and indicate that the bias is especially prevalent in shallow cumulus regimes and that may lead to increased spread in cloud feedbacks among models. Recently Vignesh et al. (2020) used COSP outputs to compare the cloud cover between CMIP5 and CMIP6 versions of 10 models, including CESM (comparing CAM4–CAM6), and Zhang et al. (2019) provide a detailed evaluation of the E3SM atmosphere model (EAM, which is closely related to CAM, and is another high-sensitivity model, Golaz et al., 2019). Both of those studies show some improvements over earlier generations of models.

This study uses COSP and satellite observations to evaluate the simulation of clouds across three generations of the CESM lineage (described in Section 2.1): CAM4 and CAM5 from previous generations and CAM6 and E3SM representing the current generation. We include E3SM because it is similar to CESM2(CAM6) but took a slightly different development path from CESM1(CAM5). This analysis provides a detailed account of the cloud performance across the models and provides an opportunistic (if incomplete) picture of how different choices in model development and tuning impact the cloud climatology. The updated observations, described in Section 2.2, provide a long-term record of radiative fluxes and cloud cover that allow robust statistical evaluation of the model performance. Spatial errors in cloud radiative effect (CRE) and cloud cover are reported in Section 3. To connect the spatial biases to the large-scale environment, we conditionally sample the cloud-related quantities in dynamical regimes which provides additional context for the biases and highlights specific regimes that should be targeted for additional, process-based evaluation. The analysis presented here indicates that clouds have improved through the model generations, but the differences among CAM5, CAM6, and E3SM are relatively small. To continue to improve the models, we suggest that future development should use the satellite observations of cloud properties and vertical structure as constraints along with the top-of-atmosphere radiative fluxes which have been tuning targets for all of these models.

2. Methods

2.1. Models and Simulations

The global atmosphere models used in this study share a common lineage and represent three distinct generations of climate model development. The older generations are represented by CAM4 and CAM5. CAM4 was used as the atmospheric component of the Community Climate System Model (version 4) which was released in 2010 (Gent et al., 2011). The parameterized physics of CAM4 was an update from the earlier CAM3 that dated back to 2004. The modeling system was renamed the Community Earth System Model (version 1) using CAM5, which was also released in 2010 (Hurrell et al., 2013). The change to CAM5 introduced many changes to the parameterized physics, including whole-cloth changes in the representations of shallow convection, cloud microphysics, radiative transfer, and the introduction of interactive, emissions-based aerosol effects.

The current generation of models is represented by two closely related models that were both developed starting from CAM5. The direct successor to CESM1(CAM5) is CESM2(CAM6) (Danabasoglu et al., 2020). The chief changes from CAM5 to CAM6 are in the representation of moist turbulence and clouds. Shallow cumulus convection and boundary layer turbulence are unified with grid-scale cloud physics using the CLUBB framework which replaced separate schemes in CAM5 (Bogenschutz et al., 2018). Other changes include updated cloud microphysics (Gettelman & Morrison, 2015), aerosol treatment (Liu et al., 2016), and changes to orographic and gravity wave drag. The E3SM, version 1 (Golaz et al., 2019) was developed independently from CESM2(-CAM6), but followed a similar development trajectory. While the models are similar, there are both structural and parametric differences that impact the simulated climate. The parameterized physics of the models are very

similar: they use the same schemes for turbulence and shallow convection (CLUBB), microphysics, and radiation. The changes to the drag parameterizations in CAM6 are not used in E3SM. There are some enhancements to the E3SM convection scheme, described by Xie et al. (2018), that are not included in CAM6. The tuning of the physics differs between the models (Rasch et al., 2019), which contributes to differences in their simulated climates.

All three versions of CAM that are presented here use the same finite volume dynamical core (Lin, 2004; Lin & Rood, 1997) on the same $0.9° \times 1.25°$ latitude-longitude grid. The number of vertical levels increases slightly with model version (26 in CAM4, 30 in CAM5, and 32 in CAM6), but the model top remains at 2 hPa (40 km) and the lowest model level is also the same at about 100 m thickness. Notable structural differences between EAM and CAM6 are that (a) EAM uses a spectral element dynamical core on a cubed-sphere grid, and (b) EAM has higher vertical resolution with 72 levels and model top near 0.1 hPa (60 km) and the lowest level thickness of about 20 m (Xie et al., 2018).

The simulations used here are "AMIP" experiments in the colloquial sense that they are forced by monthly sea-surface temperature and sea-ice extent derived from observations. Prescribing the ocean boundary condition keeps the model close to the observed climate while allowing the model's dynamics and parameterized physics to interact freely; alternatives such as nudging to the observed state provide stronger constraints for physics evaluation but inhibit feedbacks between the physics and dynamics that may impact free-running climate simulations. The CAM4 and CAM5 simulations are taken from Kay et al. (2012) and cover years 2001–2010. The CAM6 simulation follows the CMIP protocol for "AMIP" experiments and covers 1979 to 2014. This CAM6 AMIP simulation is a separate integration from the one in the CMIP6 archive, which extends back to 1950; the results presented are consistent with analysis of that experiment. The E3SM output was obtained from the CMIP6 archive; it is much longer than the protocol requests, covering 1870–2014. Three E3SM ensemble members are available in the archive, and we show results from "r1i1p1f1," but in our examination the results are not sensitive to choice of realization. The E3SM data includes ISCCP and CALIPSO simulator output, but MISR and MODIS simulator output were not available. In an attempt to optimize the comparisons between the models and observations we evaluated the impact of the sampling interval for much of the analysis presented; in general, there is little sensitivity to choosing different intervals, and we chose to use all model data from 2001 to the end of the simulation. As described below, the observational products extend beyond the simulations, and we use that extended time series wherever possible. Monthly means are used for all analyses, and we focus on comparing the models to satellite observations using COSP (Bodas-Salcedo et al., 2011); that is, the models' own cloud fraction fields are never used. The version of COSP in CAM4 and CAM5 is COSP1.3 (Kay et al., 2012). The version was later updated to COSP1.4 (Kay et al., 2016), which was used extensively, including in E3SM (Zhang et al., 2019) and CESM1 (Kay et al., 2018; Lenaerts et al., 2020; Takahashi et al., 2019); COSP1.4 introduced CALIPSO cloud phase diagnostics. The COSP version was updated to COSP2 in CAM6; COSP2 is a refactoring of the original code to increase performance, flexibility, and ease of implementation (Swales et al., 2018). COSP variants differ only in their performance and diagnostics available, not in their answers.

2.2. Satellite Observations and Reanalysis

Our observation sources are gridded data from CALIPSO, ISCCP, MISR, and MODIS. Although COSP includes simulators for CloudSat and Parasol, they are not used here. In addition, CERES EBAF is used for radiative fluxes (including CRE). The ERA5 reanalysis (Hersbach et al., 2020) is used for any meteorological fields, specifically vertical pressure velocity, and in all cases ERA5 is sampled based on the satellite data availability. As detailed next, efforts were made to use recent versions of all observational products and to obtain monthly data for as long a record as possible. Our choices appear to mirror those in Zhang et al. (2019), but we include updates to all the products (except CALIPSO, for which the same version is used) along with extending the records to near present.

2.2.1. CERES

We use monthly observations of CRE from the CERES EBAF (Edition 4.1) product (Loeb et al., 2018). This Level 3 product is on a regular $1° \times 1°$ latitude-longitude grid, and covers the period 2000-03 to 2021-12. With Edition 4, the clear-sky fluxes were changed substantially from previous versions (Loeb et al., 2018). The version used here (Edition 4.1) is a minor update from Edition 4.0 used by (Zhang et al., 2019).

2.2.2. CALIPSO

We use version 3.1.2 of the GOCCP product, which is a derived Level 3 product (Chepfer et al., 2010). Currently this data set spans 2006-06 to 2020-12. Derived from active remote sensing (lidar), this product accurately detects vertically resolved clouds within 480 m height bins (from 240 m to nearly 20 km above the surface). We follow the recommendation from the GOCCP developers to use the data on a $2° \times 2°$ latitude-longitude grid. Total, high, middle, and low cloud fractions are provided as separate fields. The lidar observations are impacted by the South Atlantic Anomaly (Noel et al., 2014); this appears to impact high cloud temporal trends, especially after 2016, but in long-term averages excluding the SAA region makes very small differences to the results presented here, so we include it for simplicity.

2.2.3. ISCCP

ISCCP provides data merged from a variety of satellites dating back to 1983. Recently the full data set has been reprocessed to produce the ISCCP-H series data, which is used here (W. B. Rossow et al., 2022; Young et al., 2018). Monthly cloud-top pressure versus cloud-optical-depth histograms were derived by averaging the 3-hourly data. The data is on a $1° \times 1°$ latitude-longitude grid and covers 1983-07 to 2017-06. Total, high, middle, and low cloud fractions were derived following the conventional definitions (e.g., Kay et al., 2012), including excluding $\tau < 0.3$. There is a substantial decreasing trend in global total cloud cover in the ISCCP data set from 1983 until the late 1990s; incomplete geostationary satellite coverage is also apparent in the spatial patterns of cloud cover during that period (see also Norris & Evan, 2015). These artifacts are no longer apparent after the late 1990s (e.g., Norris & Evan, 2015). For simplicity, and for added consistency with the simulations and other observational data, we present analysis of ISCCP using the period 2001–2016. Results that include 1983–2000 are qualitatively consistent with what is presented. To provide confidence in the climatological record while excluding the earlier years requires using the ISCCP-H series; previous studies (e.g., Kay et al., 2012; Zhang et al., 2019) were limited to the ISCCP-D series which stops in 2009.

2.2.4. MISR

The MISR data has been updated to version 7 of the algorithm described by Marchand et al. (2010) (see also Hillman et al., 2017; Hillman et al., 2018). The data used here covers 2000-03 to 2021-11. Because it uses multiple viewing angles, the MISR product is particularly well-suited for detecting cloud-top height, especially of low-level clouds. The MISR data is only available over ocean. Total, high, middle, and low cloud fractions were derived from the CTH-τ histograms in the same way as for ISCCP and MODIS. As with the other products $\tau < 0.3$ was neglected, as such tenuous clouds are not well-observed by MISR. Some retrievals fail to record a confident cloud-top height and are placed into a separate bin; these are included in the total cloud fraction, but do not contribute to high, middle, or low cloud fraction.

2.2.5. MODIS

The MODIS COSP Level-3 product, or MCD06COSP, is updated to span 2002-07 to 2022-07, as described by Pincus et al. (2022). The monthly data is used here. MODIS distinguishes partly cloudy pixels from cloudy pixels, and in previous assessments (e.g., Kay et al., 2012; Zhang et al., 2019) partly cloudy pixels were not included, so MODIS-based estimates of cloud cover were systematically lower than ISCCP (Pincus et al., 2012, 2022). The MODIS CTP-τ histograms used here combine cloudy and partly cloud pixels and normalize by the pixel count from the cloud retrieval. The MCD06COSP product combines observations from the Aqua and Terra spacecraft, and is restricted to daytime observations to be consistent with COSP output.

In general, analyses are restricted to regions equatorward of 60° latitude. This is because the passive observations become unreliable over ice-covered regions. The CALIPSO observations are considerably more reliable at higher latitudes, but we usually consider only equatorward of 60° latitude to allow comparison across products. The exception is the presentation of maps, which show results in the polar regions. To make direct, quantitative comparisons between satellite products and the models, some remapping is required. To accomplish this, we use NCO's ncremap with its native, conservative remapping algorithm (Zender, 2008). Testing revealed nearly identical results when using ESMF's conservative algorithm and somewhat larger differences when using TempestRemap's algorithm (Ullrich & Taylor, 2015). Slight differences in global averages are obtained for global mean total cloud fraction between the original and remapped fields. All the grids are of similar spatial resolution (CALIPSO being the coarsest at 2°), and sensitivity testing with different grid choices showed no noticeable

differences, so we chose to use CAM's $0.9° \times 1.25°$ grid for convenience. One noticeable impact of the remapping occurs near the sea ice edge in the MISR observations; errors in the remapped data around the ice edge can appear as artifacts in the climatology, but these artifacts have little impact on the presented results.

2.3. Dynamical Regimes and Statistical Methods

To better understand the differences across the model versions, we investigate the distribution and characteristics of clouds conditionally sampled based on the large-scale environment. In particular, we rely on the vertical pressure velocity at 500 hPa (ω_{500}) to partition dynamical regimes, separating convective and subsiding regimes (following, e.g., Bony et al., 2004).

Using monthly mean ω_{500} is appropriate over tropical oceans (Wyant et al., 2006). In the extratropics, the long averaging time tends to mix the ascending and descending regimes of cyclones, so the distribution of ω_{500} ($P(\omega_{500})$) tends to be focused around zero. It is more informative to use daily ω_{500} (e.g., Grise & Medeiros, 2016), but daily data is not available for all the simulations so we show the results from the monthly values as a broad-scale perspective on the Southern Ocean region.

In addition to the dynamical regimes decomposition, we also desire metrics for model performance, and here we are particularly concerned with evaluating the spatial distribution of cloud properties. A common measure of model error for a spatial field is the root-mean square error: $RMSE = \sqrt{N^{-1} \sum_i (X_M - X_R)^2}$. Here the M and R subscripts denote the model solution and the reference (usually observation), respectively. Usually the simple average shown is replaced with the area-weighted average appropriate for the spatial mesh. The RMSE provides a useful, concise metric for the difference, but it blends different aspects of the error. Taylor (2001) provides a deconstruction of RMSE into the "bias" (i.e., difference of the means) and the "centered pattern RMS difference" which can be expressed in terms of the variances and correlation coefficient between the two fields. That motivated a graphical presentation now known as a Taylor diagram that summarizes the statistics. A different approach to decompose the spatial error is to square the RMSE (i.e., the MSE) and normalize by the spatial variance of the reference field, that is, $NMSE = N^{-1} \sum_i (X_M - X_R)^2 / N^{-1} \sum_i \left(X_R - \overline{X_R} \right)^2$. Similar to Taylor's decomposition of the centered pattern RMS difference, normalized mean square error (NMSE) can be decomposed into three terms that involve the spatial standard deviations (σ) and the correlation coefficient between the fields (r). The components have been termed the unconditional bias, the conditional bias, and the phase error: $NMSE = U + C + P$ (Simpson et al., 2020). Briefly, the unconditional bias, $\left(\left(\overline{X_M} - \overline{X_R} \right) \sigma_R^{-1} \right)^2$, is related to the amplitude error (i.e., the difference of the means of the fields); the phase error is given by $1 - r^2$; and the conditional bias contains both amplitude and phase errors and is expressed as $\left(r - \sigma_M \sigma_R^{-1} \right)^2$. The conditional bias can be interpreted using the scaled variance ratio, $SVR = (\sigma_M / \sigma_R)^2 NMSE$; when $SVR > NMSE$ it indicates too much spatial variance in the model field (see Simpson et al., 2020, for additional discussion). The information contained in these two approaches is the same, but they provide slightly different perspectives on the errors. Taylor diagrams usually emphasize the spatial correlation and difference in the spatial variance. The NMSE, especially when visualized as a stacked bar graph, provides a more direct, non-dimensionalized comparison between the amplitude and phase errors.

The passive instrument data sets and simulator output are largely organized into joint histograms of cloud-top-pressure (or height; CTP/CTH) and cloud optical depth (τ). These histograms, pioneered in ISCCP, organize clouds along two dimensions that are directly associated with the SWCRE (τ) and LWCRE (CTP or CTH). For each grid cell (model or gridded observational product), the histograms are constructed by binning the data into a coarse CTP-τ grid. It should be noted, however, that only the detected cloud is recorded in this data structure (i.e., non-cloud is not included). This has the advantage that summing the values provides the cloud fraction, but the disadvantage is that the total is not unity so these are not true histograms (i.e., they are not approximations of the probability distribution). The monthly histograms are averages of the higher frequency data (e.g., 3-hourly for ISCCP, 1-hr radiation timesteps for the models).

The CTP-τ histograms provide more information than just the cloud fraction, but the added dimensionality can make it difficult to make holistic assessments. One common method to use more of the histogram information is to use the marginal distributions (i.e., summing over either the vertical or optical depth dimension, see e.g., Kay

et al., 2012). Another method that has made use of the histograms is to use them to cluster data into "weather states" (e.g., Tselioudis et al., 2013; Tselioudis et al., 2021). In order to perform such clustering, a distance metric between histograms must be defined. The functions that are used for that distance may themselves be useful as diagnostics and we apply one here. The most common choice of distance metric is the Euclidean distance (Wu, 2012); other options such as the relative entropy are also sometimes used. In most approaches the resulting distance provides a measure of the cumulative difference of the histograms by pairwise comparison of elements; for example, the Euclidean distance is $\sqrt{\sum_i (x_i - y_i)^2}$ where x and y are CTP-τ histograms and i is the "flattened" index. Such approaches are efficient and easy to implement, but they neglect the information encoded in the CTP-τ space.

To make more use of the two-dimensional information contained in CTP-τ histograms, we adopt an alternative metric for the distance between any two histograms: the Earth Mover's Distance (EMD), which is also often called the Wasserstein distance (Engquist & Froese, 2014; Villani, 2003). The EMD is the solution to an optimal transport problem that can be conceptually understood as providing the solution that allows the transformation of one distribution to another such that the movement of "mass" (or here probability of cloud detection) is minimized. To apply the EMD algorithm, CTP-τ histograms are normalized so they sum to unity; normalizing the individual histograms means that the EMD metric does not include the mean bias, which can be discerned anyhow by direct comparison of the total cloud fraction. To calculate the "distance" that probability must be transported, we use the simple Euclidean distance between bins of the CTP-τ histogram assuming uniform spacing (i.e., the penalty associated with transporting to an adjacent τ or CTP bin is constant). The advantage of using the EMD to compare histograms is that it penalizes slight shifts in the cloud distribution less than large shifts. The disadvantage of the EMD is that it is much more computationally demanding than simpler metrics; we make use of the Python Optimal Transport package (Flamary et al., 2021). The result of solving the optimal transport problem provides a scalar value that represents the minimal amount of probability that needs to be transported to make the two input distributions match; that is, it provides a metric for how different the distributions are with zero indicating they are identical and larger values indicating greater difference.

3. Results

Top-of-atmosphere CRE has been an explicit tuning target during development of these models with particular attention paid to the global and time average. Figure 1 shows the comparison with CERES EBAF Ed4.1, which was developed after CAM4 and CAM5 (and was not used during CAM6 development). In general, CAM6 shows improvements compared to the earlier versions, though CAM4's LWCRE is remarkably successful in terms of global mean bias and RMSE. Although they are similar models, E3SM shows slightly larger shortwave bias and RMSE than CAM6, and slightly smaller longwave errors. Some additional context of these spatial errors can be gleaned with the NMSE (defined in Section 2), which can be decomposed into unconditional, conditional, and phase error contributions. Figure 2 shows the NMSE by component for SWCRE and LWCRE for all four models. The total error, given by the NMSE, is smallest in E3SM for both SWCRE and LWCRE. Figure 2 uses data equatorward of 60° latitude; using the full global domain results in CAM6 having a slightly smaller SWCRE NMSE than E3SM, but in both cases the values are quite close between the models. For both CAM6 and E3SM, the improvement from CAM5 is mainly from reducing the unconditional bias in the longwave while the shortwave improves chiefly by improving the phase error. The phase error is largest in CAM4 showing that even though global mean bias is small, the regional features are erroneous.

While the TOA CRE is a crucial quantity for climate models to capture, it is not a sufficient evaluation of the quality of the cloud climatology in a model. Both cloud amount and cloud albedo impact SWCRE, for example, so it is possible to match observations of SWCRE with unrealistic cloud physical properties; this situation is common in models and leads to the "too-few-too-bright" bias (Nam et al., 2012). With relatively long observational records and the use of satellite simulators to facilitate comparison, several aspects of cloud cover can be evaluated with reasonable confidence. Using both CRE and cloud cover provides a stronger constraint on a model's cloud climatology than either provides separately. This combination is what we demonstrate in the rest of this section with an evaluation of the cloud cover climatology.

Figure 3 shows the climatology of cloud cover for ISCCP, MISR, and CALIPSO in the top row. Because of the many differences in these data sets (Pincus et al., 2012), the cloud cover differs across the observations. These
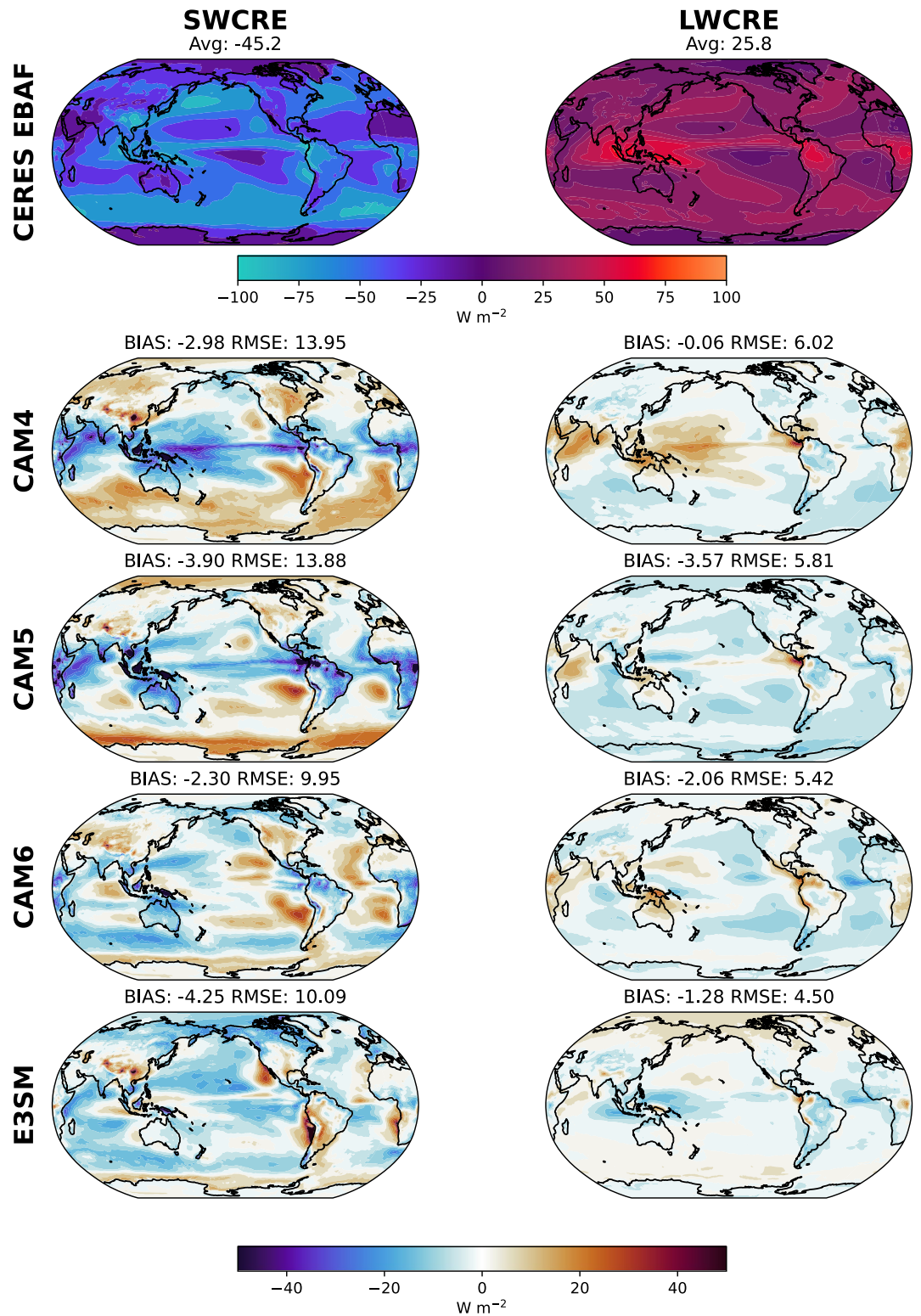
Figure 1. (top) CERES EBAF observations of (left) SWCRE and (right) LWCRE. The area-weighted global average is shown at the top of each. (lower rows) Differences between model and observations for each model version. The global mean difference (bias) and the RMSE are noted at the top of each panel.
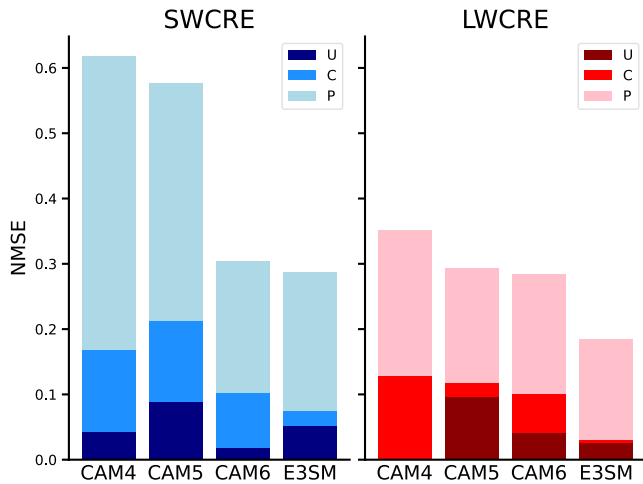
Figure 2. Normalized mean squared error for (left) SWCRE and (right) LWCRE between each model version and CERES EBAF for the region 60°S–60°N. The bars are shaded to show contributions from unconditional bias (U), conditional bias (C), and phase error (P) as defined in the text.

differences (of a few percent globally), however, are small compared to the difference between the observations and models. CAM4 markedly underestimates cloud cover, while CAM5, CAM6, and E3SM appear to be improved but still underestimate global cloud cover by more than 10% compared to the passive remote sensing products (ISCCP, MISR) and 7%–10% compared to CALIPSO. The values provided by Figure 3 are global; excluding polar regions produces qualitatively similar comparisons among observations and models.

Figure 4 shows the NMSE for the total cloud cover (60°S–60°N). The improvement from CAM4 to later versions is largely due to the large reduction in the unconditional error, essentially the bias, as shown by the black-shaded part of each bar. The results are mixed when comparing the newer models, but across the satellite products the conditional error (medium gray shading) is larger in CAM6 than either CAM5 or E3SM. The conditional error arises from both phase and amplitude errors, but when the scaled variance ratio (red dots in the figure, defined in Section 2) is larger than the NMSE the conditional error is mainly from spatial variance. All the newer models show this to be the case and therefore show excessive spatial variance, and CAM6 has the largest SVR.

The maps of Figure 3 suggest that CAM6's and E3SM's excessive spatial variability comes from a large increase in cloud cover at high latitudes, especially the Southern Ocean. With respect to ISCCP, the current-generation CAM6 and E3SM show very different errors in the tropical east Pacific: CAM6 shows much too little cloud cover in the stratocumulus regions while E3SM shows excessive cloud cover along the equator from South America into the central Pacific. That is corroborated by the zonal mean view in Figure 5 comparing the CALIPSO observations to the models for total, high, mid-level, and low clouds. At high latitudes, CAM6 and E3SM display strong positive biases in low-level clouds. All the models have too little cloud cover at low latitudes, and this predominantly arises from too little low-level cloud. The overestimation of high-latitude cloud and underestimation of low-latitude cloud causes the spatial variation to be larger than observed, especially apparent for CAM6, leading to the conditional bias shown in Figure 4. Similar results are found using the other observational data sets.

3.1. Histogram Error

So far we have presented diagnostics that illustrate the spatial errors in the model clouds using standard measures of cloud cover and CRE. The satellite products and simulators for ISCCP, MODIS, and MISR provide additional information in the form of the CTP-τ histograms. The total cloud cover and layer cloud amounts are derived by summing over the histogram. As described in Section 2, we can use measures of similarity between the histograms to make an additional assessment of model errors.

Figure 6 shows the EMD metric calculated using the climatology from the satellite products and model simulations. The maps show the EMD at all available locations, but the global average (shown in gray) is restricted to ±60° latitude for all three products. As discussed next, there appears to be decreasing histogram error through model generations. Errors for MISR (rightmost column) are systematically larger than for ISCCP and MODIS, but this is a result of the larger histogram grid for MISR (6×16 as opposed to 6×7) and the simple Euclidean cost function that is used in the EMD calculation.

The global average histogram errors decrease through model generations. The improvement is striking between CAM4 and the others, while CAM5 and CAM6 show more subtle differences from each other. Only ISCCP data was available for E3SM, and it shows the lowest global average EMD. Comparing CAM6 and E3SM, the E3SM EMD is smaller in the Southern Ocean and eastern ocean basins while it is larger over much of the tropical oceans. All the models, in all three satellite perspectives, show errors associated with trade wind regions. The subtropical stratocumulus regions show somewhat smaller errors than the trade wind regions, but do not appear to be reduced much through the model generations. Isolating these regimes can be more objective using conditional sampling, as discussed next.
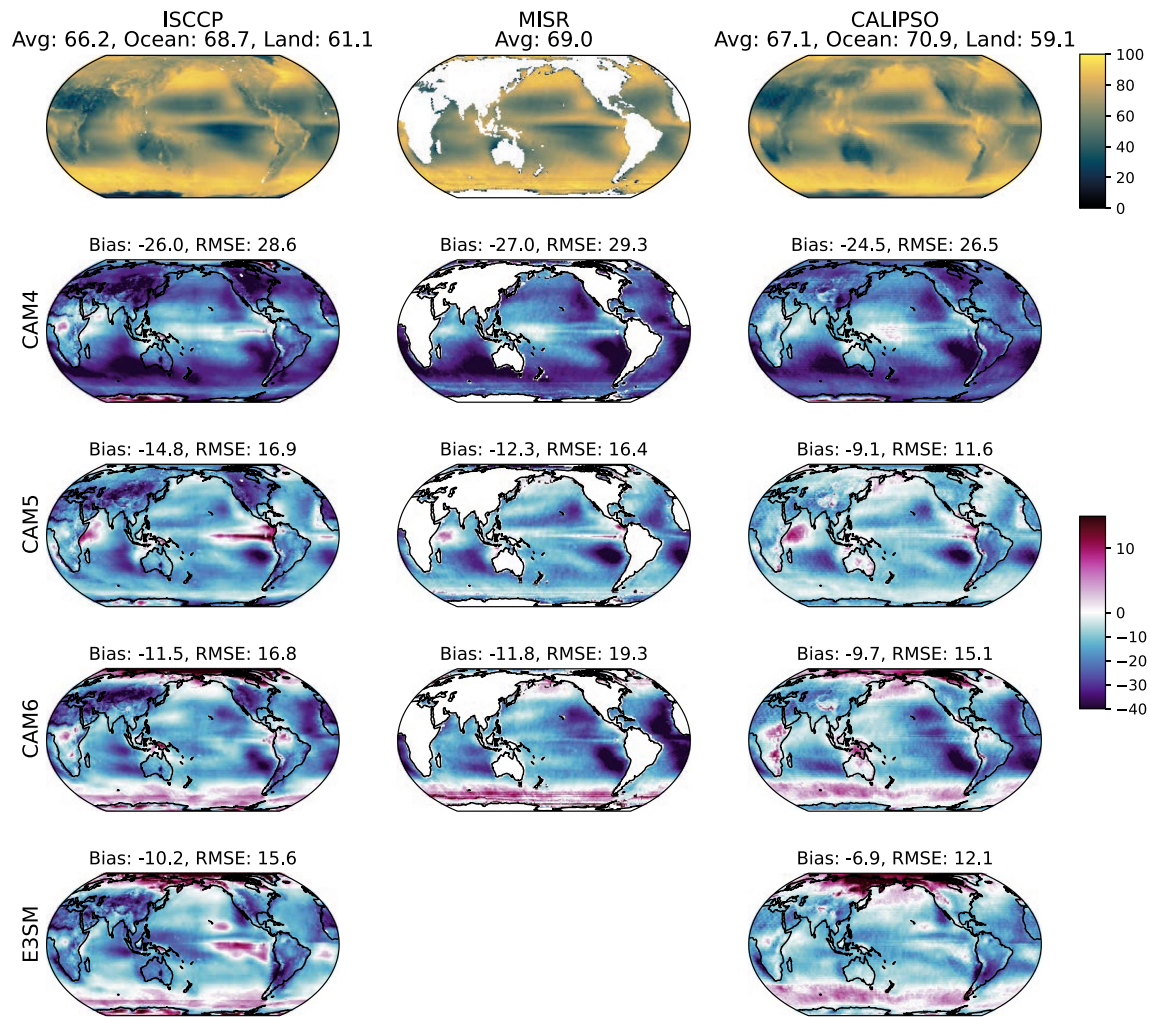
Figure 3. (top) Total cloud cover (percent) from (left to right) ISCCP, MISR, and CALIPSO. (lower rows) Differences in total cloud cover between the models and the observations. Global average values (top row) and bias and RMSE (lower rows) use the global domain and are area-weighted.
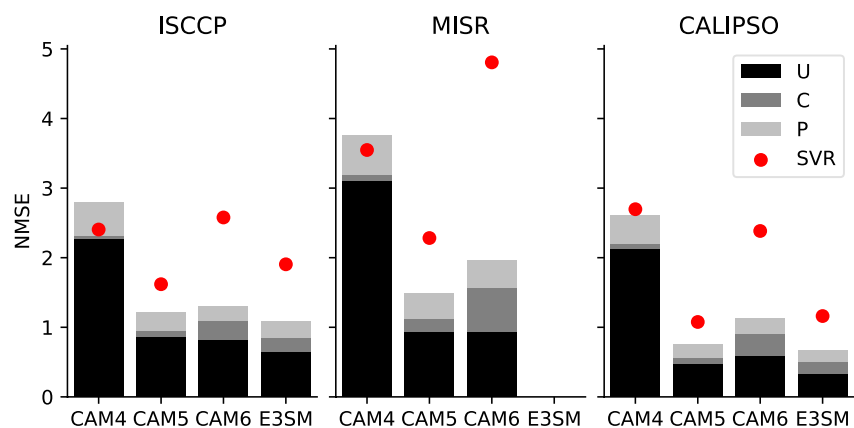

Figure 4. Total cloud cover NMSE by satellite product. Red dots show the scaled variance ratio. The calculation uses 60°S–60°N.
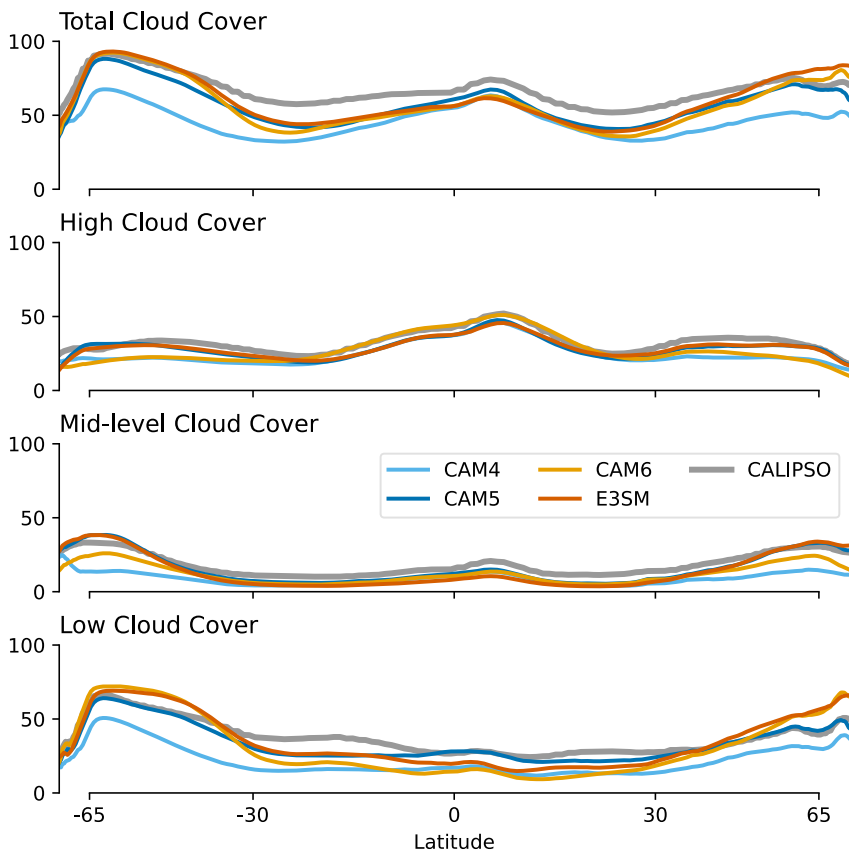
Figure 5. Zonal mean cloud cover from CALIPSO and the CALIPSO simulators: (top to bottom) total, high-, middle-, and low-level cloud cover (percent).
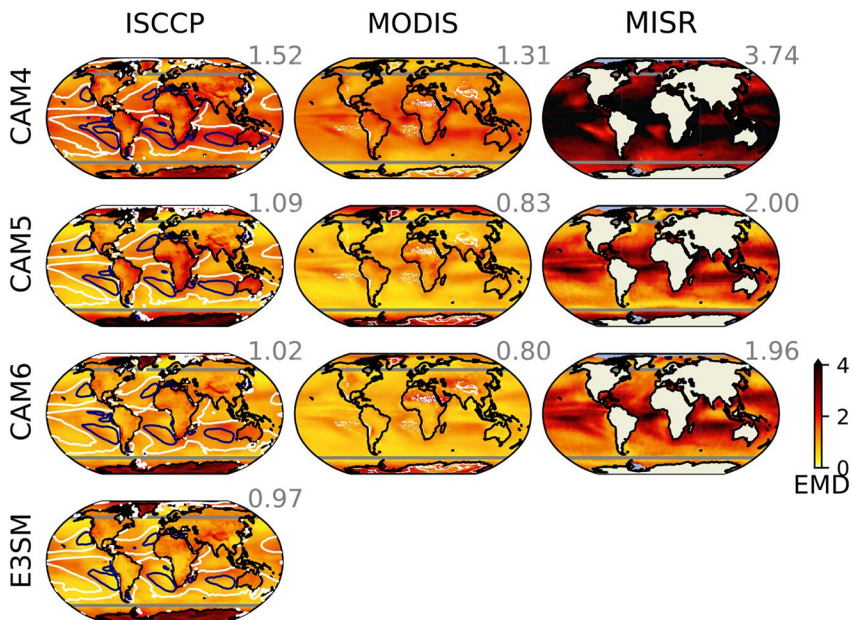

Figure 6. Histogram distance metric, EMD, for the climatology of each model. Averages for ±60° latitude are shown in the upper right of each panel (gray lines are included on the maps to note the region for the average). Contours in the left column delineate $\omega_{500} = 0$ (white) and the 75th percentile of ω_{500} (navy).
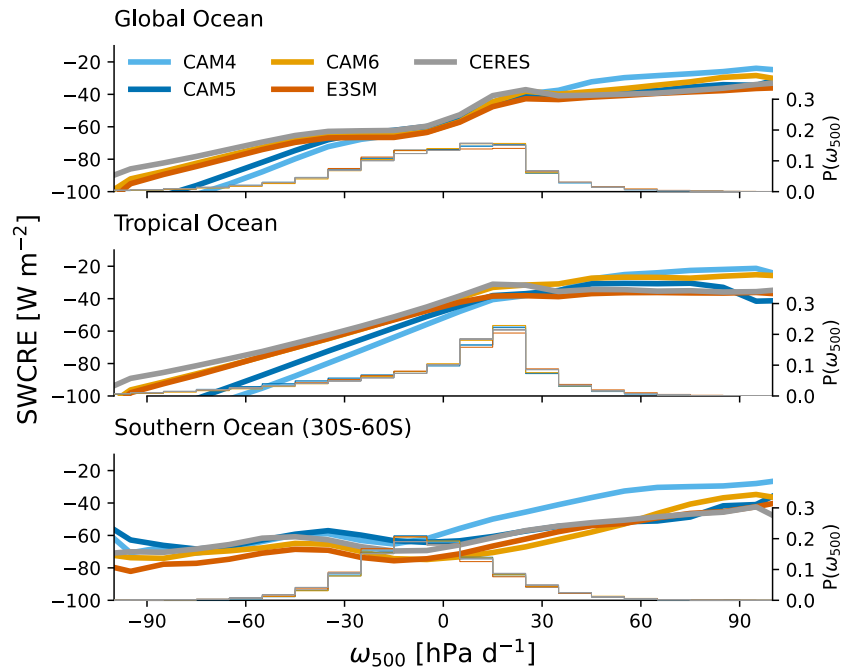
Figure 7. SWCRE binned by ω_{500} for the (top to bottom) global ocean (60°S–60°N), tropical ocean (30°S–30°N), and southern ocean (30°S–60°N). The histogram of ω_{500} is shown by thin lines (axis on the right).

3.2. Dynamical Regimes

In this section, the COSP-derived cloud cover is cast in the context of monthly dynamical regimes based on the mid-tropospheric vertical pressure velocity (ω_{500}). Mid-tropospheric vertical velocity differentiates dynamical regimes by separating regions with net ascent (indicating convectively active regions) from those with net descent (aka subsidence) that have a drier free troposphere and often foster cloud-topped boundary layers (Bony et al., 2004). As mentioned in Section 2, at monthly timescales, ω_{500} does reasonably well at partitioning regimes in the tropics, but in the extratropics states with weak ω_{500} are overemphasized. We also concentrate attention on ocean locations to avoid the additional complications that are induced by land-atmosphere interactions and topographic effects.

Figure 7 shows the SWCRE partitioned by ω_{500} for all ocean locations (top) as well as tropical oceans (middle) and the Southern Ocean (bottom) separately. The histogram of ω_{500} is also shown (thin lines). Globally, the models show overall good agreement with the CERES observations (binned using ERA5 ω_{500}). CAM6 and E3SM are noticeably improved in strong convective regimes, and CAM4 is farthest from observations in most regimes. This global improvement in CAM6 and E3SM appears to be mainly from tropical oceans (middle panel). The models appear relatively close to the observations in the subsidence regimes. Within the small differences, CAM6 is closest to the CERES observations in weak subsidence regimes, and E3SM agrees very well with observations in the strong subsidence regimes ($\omega_{500} > 25$ hPa d^{-1}). In the Southern Ocean, the models capture the shape of the SWCRE(ω_{500}) distribution with CAM6 and E3SM showing a bias toward too much reflection in subsidence regimes while CAM4 and CAM5 have SWCRE that is too weak in most regimes. The large-scale circulation, measured by the histogram of ω_{500}, is well captured by all the models; partly that should be attributed to these simulations being forced by observed SST and sea ice.

Figure 8 shows the vertical structure of clouds in dynamical regimes as observed by CALIPSO (top row). As would be expected, convective regimes tend to have more high-level cloud while subsidence regimes are dominated by low-level cloud. It is notable that the tropical convective regimes show some indication of a tri-modal structure with enhanced cloud fraction below about 2 km, around 7 km, and around 13 km (nb. Johnson et al., 1999). The Southern Ocean has more uniform cloud cover in monthly averages because the region is dominated by eastward propagating cyclones that mix cloud states on synoptic timescales. Nevertheless, the general tendency for more
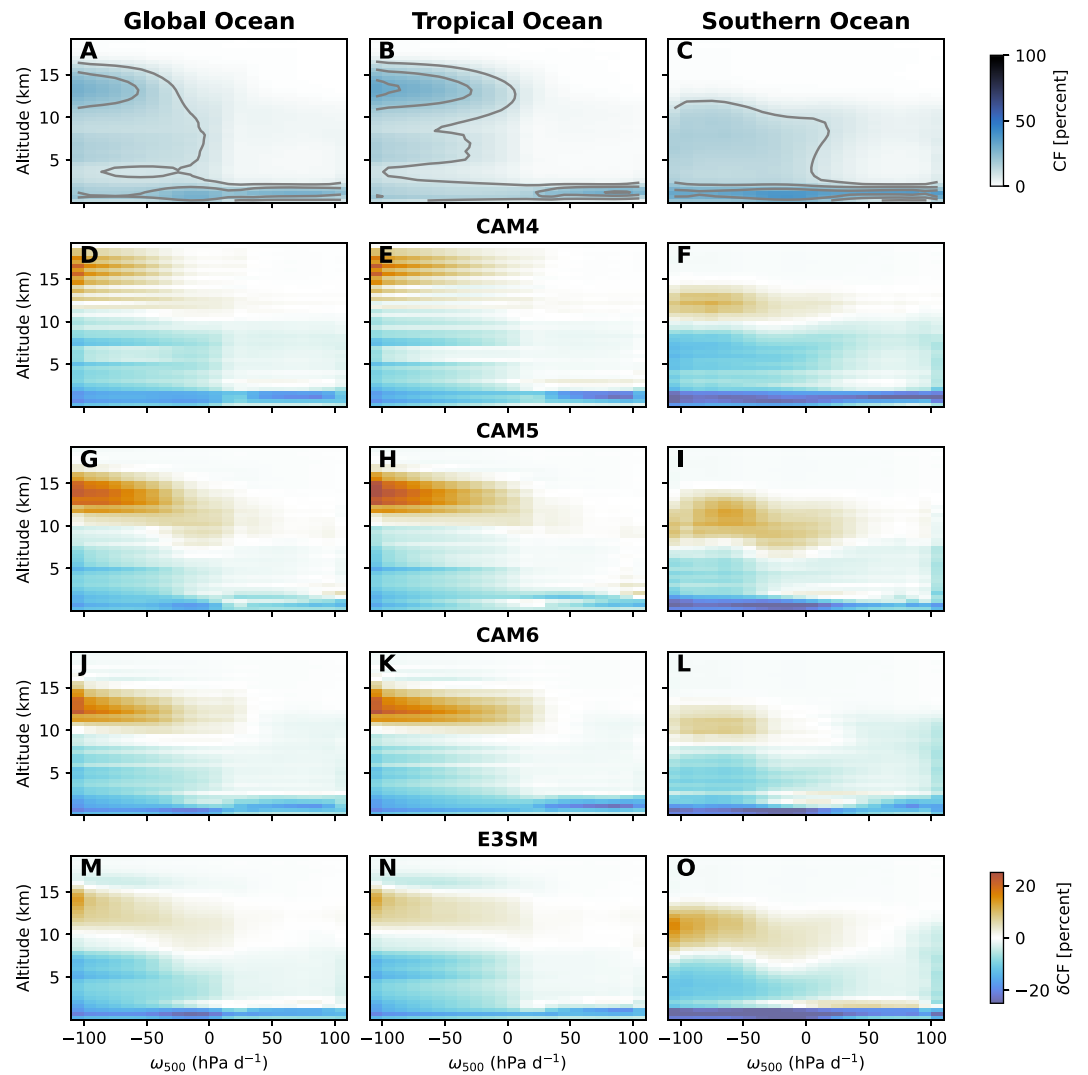
Figure 8. CALIPSO cloud fraction binned by ω_{500}. (top row) The observed cloud fraction, and (lower rows) the difference of the observations from each model. Only ocean locations are used, and the global region here is 60°S–60°N; the tropical oceans are defined as 30°S–30°N., and the Southern Ocean is 60°S–30°S.

cloud at higher levels in convective regimes is apparent, as is a pronounced peak in cloud fraction at low-levels in the very common weak convection and weak subsidence regimes.

The lower rows of Figure 8 show the difference between the observed and simulated CALIPSO clouds. In every case, the models show excessive upper-troposphere cloud fraction in strong convective regimes. The height where the bias occurs decreases with model generation. The bias becomes smaller in weak convection regimes, but remains positive into subsidence regimes indicating that the models have too much cirrus cloud in most conditions. CAM4 appears to have some linear features in its bias, but we have not determined a reason for such a pattern. The models also all show a deficit of low-level clouds. Especially relevant for the global energy balance, the subsidence regimes over tropical oceans show significant errors, as also suggested in earlier figures.

Biases in the vertical structure of clouds shown in Figure 8 result in biases in the total cloud cover. The CALIPSO-based total cloud cover is shown in dynamical regimes in Figure 9. Despite the excess cloud cover at upper levels in convective regimes, the models slightly underestimate total cloud fraction in these regimes. Over tropical oceans, the models significantly underestimate cloud cover in weak convection and subsidence regimes. The model bias is smaller in the southern ocean region in the weak convection and weak subsidence regimes
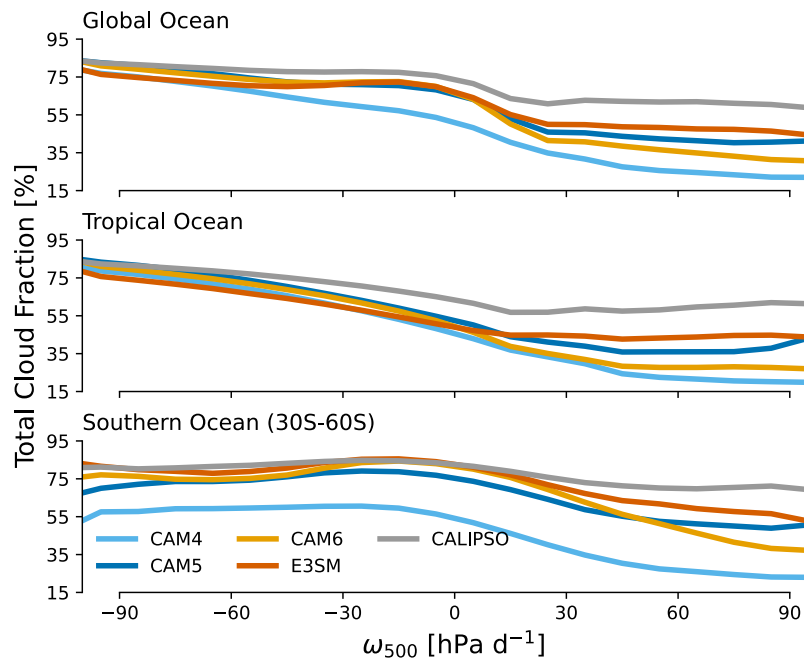
Figure 9. CALIPSO total cloud fraction binned by ω_{500}: (top) the global oceans (60°S–60°N), (middle) tropical oceans (30°S–30°N), and (bottom) the Southern Ocean (60°S–30°S).

(except for CAM4); the magnitude of the bias is somewhat sensitive to choice of latitudes, as could be inferred by the maps in Figure 3.

The EMD metric provides a concise way to characterize the CTP-τ histogram errors by dynamical regimes. Figure 10 shows this for all models compared against ISCCP. For these comparisons, EMD is calculated between
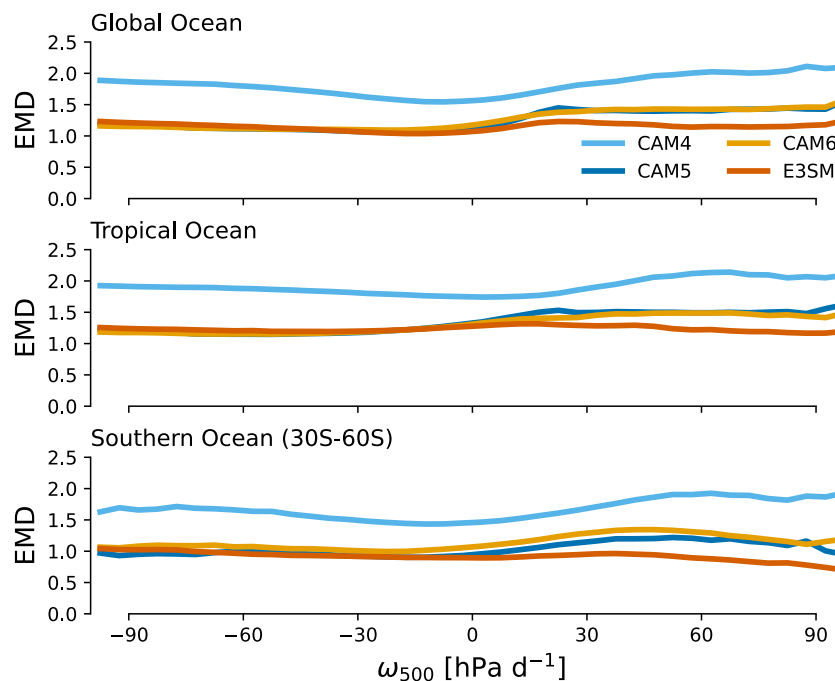

Figure 10. Average EMD for monthly ISCCP histograms binned by the model's ω_{500}. The EMD is calculated at each grid cell for each month that occurs in both the model and observational data.
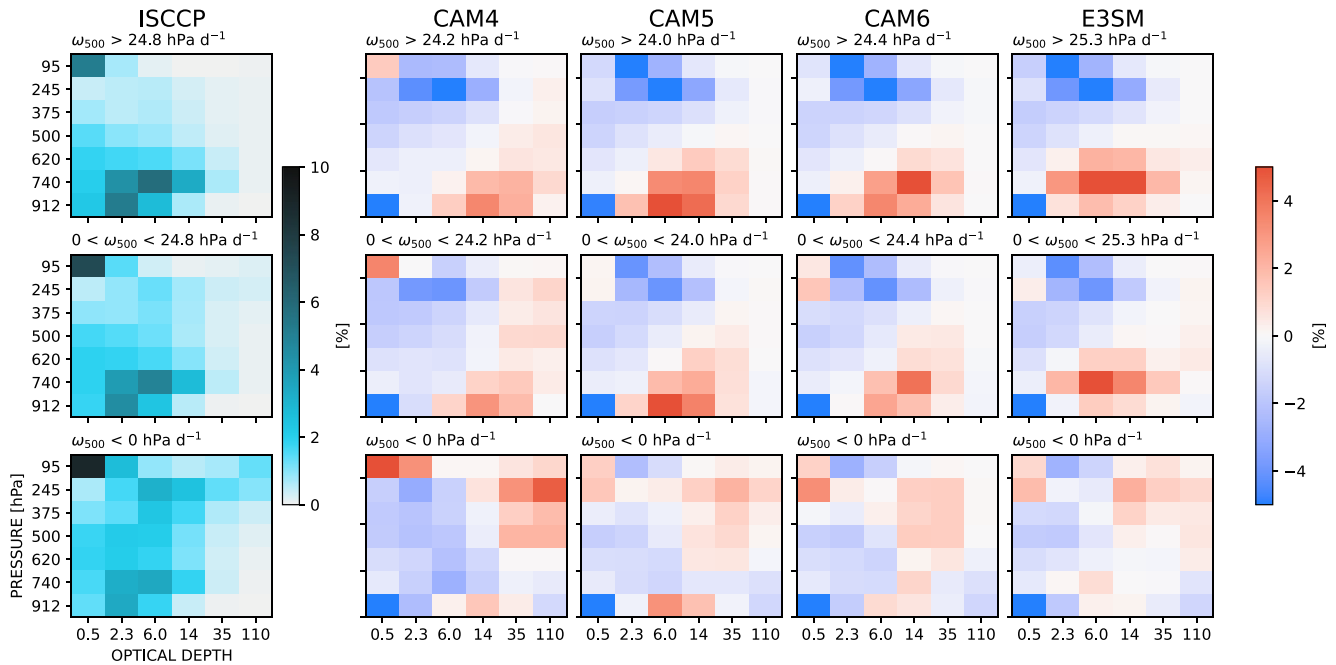
Figure 11. CTP-τ histograms from ISCCP (left column) over tropical oceans, divided into strong subsidence (top), weak subsidence (middle), and convective (bottom) regimes. Strong subsidence is delineated by the 75th percentile of the ω_{500} distribution. The columns to the right show the model minus observation for the same sampling; the value of the ω_{500} 75th percentile is reported in the title.

the satellite and model data for each month that is available for both (unlike Figure 6 that used the climatological average); this provides better sampling across regimes, but introduces more noise associated with internal variability. Those EMD values are averaged within bins of the model's vertical velocity field. Using EMD as a measure of cloud structural error with these dynamical regimes indicates that CAM4 is an outlier, having much larger errors across all regimes than the other models. The other models all show similar errors in convective regimes, but diverge in the stronger subsidence regimes where E3SM shows smaller errors than CAM5 and CAM6.

To investigate the differences among the CTP-τ histograms that lead to the differences in Figure 10, composite histograms are constructed for the tropical ocean using ω_{500}. Figure 11 shows the observed ISCCP histograms for convective regimes ($\omega_{500} < 0$ hPa d^{-1}), weak subsidence regimes ($0 < \omega_{500} < \omega_{500}^{(75)}$ hPa d^{-1}), and strong subsidence regimes ($\omega_{500} > \omega_{500}^{(75)}$). Here $\omega_{500}^{(75)}$ is the 75th percentile of the monthly ω_{500} distribution, and values are given on figure. These regimes are delineated in the climatology in the left column of Figure 10; the strong subsidence regime is focused in the eastern ocean basins where subtropical stratocumulus are expected. The composite histograms of Figure 11 suggest the models tend to have clouds that are too optically thick, especially in subsidence regimes. The lower EMD for E3SM in the strong subsidence regimes appears to be due to the clouds being slightly optically thinner with cloud-top pressure better matching observations while all versions of CAM have too much cloud in the lowest pressure bin.

4. Discussion

This evaluation of cloud cover in CAM and E3SM in simulations forced by observed SST and sea ice provides additional context for interpreting the high climate sensitivity of the models. The prescribed SST and ice ensure a relatively realistic large-scale circulation, and the PDF of ω_{500} in Figure 7 confirms that the distribution of dynamical regimes is close to that of ERA5. Within these dynamical regimes, each atmosphere model exhibits some error in cloud properties. By most measures, CAM5, CAM6, and E3SM are much closer to observations than CAM4. It is difficult to definitively state whether the clouds in any of the more recent models is superior, as each have relative strengths and weaknesses, but E3SM tends to show slightly better agreement with observations than CAM.

The weakening of SWCRE in the Southern Ocean under greenhouse gas forcing has been implicated in the high climate sensitivity of recent models (e.g., Zelinka et al., 2020). The SWCRE bias in the Southern Ocean in CAM6 and E3SM differs from CAM5 (Figure 1), with negative bias in the southern midlatitudes indicating more reflected shortwave radiation than CERES observations. This bias is connected to excessive cloud cover, as shown in Figure 3. Compared to CALIPSO observations, it appears the excessive cloud cover is associated with too much low-level cloud (Figure 5), and that finding is corroborated by the errors being consistent with the biases in subsidence regimes in the Southern Ocean (Figures 7 and 8). The histogram errors in the Southern Ocean (Figure 10), indicate that the cloud characteristics in E3SM are closer to ISCCP than CAM5 or CAM6. Indeed, comparing the composite histograms (not shown, but similar to Figure 11) indicates they have similar structure, but CAM6 and E3SM have more cloud cover than CAM5 and a bias toward higher τ. Previous discussions of the impact of the Southern Ocean region on the high climate sensitivity of CESM2 (e.g., Bjordal et al., 2020) have emphasized the relative increase in supercooled water and decrease in cloud ice compared to the previous generation CESM1(CAM5); starting from a base state with more liquid reduces the negative cloud phase feedback. The liquid and ice water content in CESM2(CAM6) and E3SM have been reported to be more consistent with observations than CESM1(CAM5) (e.g., Yang et al., 2021). One might argue, therefore, that starting from a present day climate with improved mixed-phase clouds in the Southern Ocean, CESM2(CAM6) and E3SM reduce a negative cloud phase feedback that is unrealistically strong in CESM1(CAM5). Our results may suggest that, separate from the cloud phase feedback, the excessive cloud fraction along with the positive τ bias could also contribute to the high climate sensitivity, especially as the Southern Ocean loses substantial cloud cover as warming increases (cf., Bacmeister et al., 2020).

The reduced cloud phase feedback in CESM2(CAM6) and E3SM allows the loss of low-level subtropical clouds to strongly contribute to the high climate sensitivity (Bacmeister et al., 2020; Golaz et al., 2019). These regions have too little cloud cover in all the models (Figures 3 and 8). The histogram error shown in Figure 10 suggests the cloud structure in subsidence regimes is best in E3SM. In strong subsidence regimes, E3SM also better captures the SWCRE (Figure 7). Taken together, this indicates that E3SM is providing a superior representation of clouds in tropical subsidence regimes, but is underestimating the cloud cover in the weak subsidence regimes (i.e., the regimes that carry the most statistical weight). In tropical, weak-subsidence regimes, CAM6 has SWCRE consistent with CERES, but also has too little cloud cover. Figure 11 shows that all the models have clouds that are too optically thick. Combined with their bias toward too little cloud cover, it indicates that these models exhibit the longstanding "too-few-too-bright" bias (Nam et al., 2012). As the CAM6 bias appears to affect all tropical subsidence regimes, it appears that both stratocumulus and shallow cumulus regimes in CAM6 are afflicted with too little cloud cover that is partially compensated by excessive reflection.

Based on these results, an objective for future model development should be to increase subtropical cloud cover while preserving (or improving) the match with CERES EBAF CRE. The cloud cover errors appear especially focused in the eastern oceans (Figure 3), possibly providing a clue that the underestimation of clouds is connected with the parameterized dynamics of subtropical stratocumulus clouds. In CAM5, the underestimation of stratocumulus was suggested to be due to the cloud layer becoming decoupled from the subcloud layer too easily (Medeiros et al., 2012); that mechanism is consistent with the errors seen in CAM6, so should be investigated. The errors in the weak subsidence regimes, shown in Figures 7–10, carry a large amount of weight because they cover vast areas of the tropical oceans. Improving the representation of shallow cumulus clouds, both in terms of cloud cover and radiative effects, also holds the potential to have a substantial impact on the models' climate sensitivity. Parameterizing these boundary layer clouds is notoriously challenging, and despite progress (e.g., Kawai & Shige, 2020; Randall, 2013), there remain difficulties in capturing interactions among parameterized processes. There are also numerous small-scale processes that may be important that are not included in current parameterizations (e.g., Zeng & Li, 2023).

This study provides examples of several methods that help to expose errors in the model cloud climatologies. The normalized mean squared error (NMSE) provides a simple method to decompose the climatological error into mean bias and spatial errors, and is complementary to the information in Taylor diagrams. Using dynamical regimes based on ω_{500} separates convective and subsidence regimes, allowing an evaluation that does not depend on choosing geographical regions. The EMD provides a concise way to investigate errors in cloud properties. These methods can be extended to seasonal analysis and aspects of climate variability, and can be elaborated upon with additional information (e.g., by also including measures of inversion strength to better separate low-level cloud regimes). Use of daily data would also provide a more detailed view, especially in the extratropics. These

diagnostics can provide guidance toward targeted analysis and experimentation with particular cloud regimes to determine the physical causes of the discrepancies with observations.

Satellite simulators allow a direct, fair comparison between simulated cloud cover and observations. Satellite records of cloud cover from multiple platforms are now reaching (or already exceed) 20 years duration, and provide a compelling record of cloud cover over the tropics and midlatitudes. With the combination of satellite simulators and long-term, high-quality observations, cloud cover can be used as a measure of climate model fidelity. Together, CRE and cloud cover organized by pressure and optical depth provide relatively strong constraints on cloud properties in the current climate. As shown here, global atmosphere models have improved over the past couple of decades, but still exhibit significant errors in cloud cover. Improving cloud cover while preserving, or improving, CRE should be a development goal. Other models likely have similarly large errors, and a multi-model evaluation should be conducted to understand the spread in cloud errors. Reducing these errors in the current climate is likely to have a discernible impact on the spread in estimates of climate sensitivity.

Appendix A: Abbreviations

Table A1 provides a list of all abbreviations used throughout the text.

Table A1
Abbreviations Used in the Text

Abbreviated	Expanded
CALIPSO	Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation
CAM	Community Atmosphere Model
CESM	Community Earth System Model
CERES EBAF	Clouds and the Earth's Radiant Energy System, Energy Balanced and Filled
CMIP	Coupled Model Intercomparison Project
CFMIP	Cloud-Feedback Model Intercomparison Project
CLUBB	Cloud layers unified by binormals
COSP	CFMIP Observation Simulator Package
CMIP	Coupled model intercomparison project
E3SM	Energy Exascale Earth System Model
EAM	E3SM Atmosphere Model
EMD	Earth Mover's Distance
ESGF	Earth system grid federation
ESMF	Earth System Modeling Framework
ERA5	ECMWF reanalysis, generation 5
ISCCP	International Satellite Cloud Climatology Project
LWCRE	longwave cloud radiative effect
MISR	Multi-angle Imaging SpectroRadiometer
MODIS	Moderate Resolution Imaging Spectroradiometer
NCO	netCDF Operators
NMSE	Normalized mean squared error
RMSE	Root-mean squared error
SAA	South Atlantic Anomaly
SWCRE	Shortwave cloud radiative effect

Data Availability Statement

The CALIPSO GOCCP (v3.1.2) data was obtained directly from the project website (GOCCP, 2023). ERA5 monthly averages are available from the Copernicus Climate Change Service (2019). ISCCP-H data were obtained in netCDF files on a 1° × 1° latitude-longitude grid from W. Rossow et al. (2016). MISR data was obtained as netCDF files from Marchand (2023). The MODIS COSP Level-3 product, MCD06COSP, was downloaded as monthly netCDF files from the data archive of the NASA Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC) (MODIS Atmosphere Science Team, 2022). The E3SM simulation data is available as part of the CMIP6 archive via the ESGF (Bader et al., 2019). The CAM simulation data used in this is available in a zenodo archive (Medeiros et al., 2023). The contents include the COSP outputs and ω_{500} as monthly means. Analysis code used to produce all figures is available in a zenodo archive (Medeiros, 2023).

References

Bacmeister, J. T., Hannay, C., Medeiros, B., Gettelman, A., Neale, R., Fredriksen, H. B., et al. (2020). CO_2 increase experiments using the Community Earth System Model (CESM): Relationship to climate sensitivity and comparison of CESM1 to CESM2. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002120. https://doi.org/10.1029/2020MS002120

Bader, D. C., Leung, R., Taylor, M., & McCoy, R. B. (2019). *E3SM-Project E3SM1.0 model output prepared for CMIP6 CMIP AMIP*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4492

Bjordal, J., Storelvmo, T., Alterskjær, K., & Carlsen, T. (2020). Equilibrium climate sensitivity above 5°C plausible due to state-dependent cloud feedback. *Nature Geoscience*, *13*(11), 718–721. https://doi.org/10.1038/s41561-020-00649-1

Bodas-Salcedo, A., Webb, M. J., Bony, S., Chepfer, H., Dufresne, J. L., Klein, S. A., et al. (2011). COSP: Satellite simulation software for model assessment. *Bulletin of the American Meteorological Society*, *92*(8), 1023–1043. https://doi.org/10.1175/2011BAMS2856.1

Bogenschutz, P. A., Gettelman, A., Hannay, C., Larson, V. E., Neale, R. B., Craig, C., & Chen, C.-C. (2018). The path to CAM6: Coupled simulations with CAM5.4 and CAM5.5. *Geoscientific Model Development*, *11*(1), 235–255. https://doi.org/10.5194/gmd-11-235-2018

Bony, S., Dufresne, J.-L., Treut, H. L., Morcrette, J.-J., & Senior, C. (2004). On dynamic and thermodynamic components of cloud changes. *Climate Dynamics*, *22*(2–3), 71–86. https://doi.org/10.1007/s00382-003-0369-6

Chepfer, H., Bony, S., Winker, D., Cesana, G., Dufresne, J. L., Minnis, P., et al. (2010). The GCM-Oriented CALIPSO cloud product (CALIPSO-GOCCP). *Journal of Geophysical Research*, *115*(D4), D00H16. https://doi.org/10.1029/2009jd012251

Copernicus Climate Change Service. (2019). ERA5 monthly averaged data on pressure levels from 1979 to present [Dataset]. ECMWF. https://doi.org/10.24381/CDS.6860A573

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., et al. (2020). The Community Earth System Model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. https://doi.org/10.1029/2019MS001916

Engquist, B., & Froese, B. D. (2014). Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, *12*(5), 979–988. https://doi.org/10.4310/CMS.2014.v12.n5.a7

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., et al. (2021). POT: Python Optimal Transport. *Journal of Machine Learning Research*, *22*(78), 1–8. Retrieved from http://jmlr.org/papers/v22/20-451.html

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The Community Climate System Model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011jcli4083.1

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., et al. (2019). High climate sensitivity in the Community Earth System Model version 2 (CESM2). *Geophysical Research Letters*, *46*(14), 8329–8337. https://doi.org/10.1029/2019GL083978

Gettelman, A., Kay, J. E., & Shell, K. M. (2012). The evolution of climate sensitivity and climate feedbacks in the community atmosphere model. *Journal of Climate*, *25*(5), 1453–1469. https://doi.org/10.1175/JCLI-D-11-00197.1

Gettelman, A., & Morrison, H. (2015). Advanced two-moment bulk microphysics for global models. Part I: Off-line tests and comparison with other schemes. *Journal of Climate*, *28*(3), 1268–1287. https://doi.org/10.1175/jcli-d-14-00102.1

GOCCP. (2023). GCM-oriented CALIPSO cloud product [Dataset]. Retrieved from https://climserv.ipsl.polytechnique.fr/cfmip-obs/Calipso_goccp.html

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., et al. (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. https://doi.org/10.1029/2018ms001603

Grise, K. M., & Medeiros, B. (2016). Understanding the varied influence of midlatitude jet position on clouds and cloud radiative effects in observations and global climate models. *Journal of Climate*, *29*(24), 9005–9025. https://doi.org/10.1175/jcli-d-16-0295.1

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M. (2022). *Climate simulations: Recognize the "hot model" problem*. Nature Publishing Group. https://doi.org/10.1038/d41586-022-01192-2

He, H., Soden, B., & Kramer, R. J. (2022). On the prevalence of high climate sensitivity models. *Authorea*. https://doi.org/10.1002/essoar.10512532.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hillman, B. R., Marchand, R. T., & Ackerman, T. P. (2018). Sensitivities of simulated satellite views of clouds to subgrid-scale overlap and condensate heterogeneity. *Journal of Geophysical Research: Atmospheres*, *123*(14), 7506–7529. https://doi.org/10.1029/2017jd027680

Hillman, B. R., Marchand, R. T., Ackerman, T. P., Mace, G. G., & Benson, S. (2017). Assessing the accuracy of MISR and MISR-simulated cloud top heights using CloudSat- and CALIPSO-retrieved hydrometeor profiles. *Journal of Geophysical Research: Atmospheres*, *122*(5), 2878–2897. https://doi.org/10.1002/2016jd025510

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., et al. (2013). The Community Earth System Model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, *94*(9), 1339–1360. https://doi.org/10.1175/bams-d-12-00121

Johnson, R. H., Rickenbach, T. M., Rutledge, S. A., Ciesielski, P. E., & Schubert, W. H. (1999). Trimodal characteristics of tropical convection. *Journal of Climate*, *12*(8), 2397–2418. https://doi.org/10.1175/1520-0442(1999)012<2397:tcotc>2.0.co;2

Kawai, H., & Shige, S. (2020). Marine low clouds and their parameterization in climate models. *Journal of the Meteorological Society of Japan. Ser. II*, *98*(6), 1097–1127. https://doi.org/10.2151/jmsj.2020-059

Kay, J. E., Bourdages, L., Miller, N. B., Morrison, A., Yettella, V., Chepfer, H., & Eaton, B. (2016). Evaluating and improving cloud phase in the community atmosphere model version 5 using spaceborne lidar observations. *Journal of Geophysical Research: Atmospheres*, *121*(8), 4162–4176. https://doi.org/10.1002/2015jd024699

Kay, J. E., Hillman, B. R., Klein, S. A., Zhang, Y., Medeiros, B., Pincus, R., et al. (2012). Exposing global cloud biases in the Community Atmosphere Model (CAM) using satellite observations and their corresponding instrument simulators. *Journal of Climate*, *25*(15), 5190–5207. https://doi.org/10.1175/JCLI-D-11-00469.1

Kay, J. E., L'Ecuyer, T., Pendergrass, A., Chepfer, H., Guzman, R., & Yettella, V. (2018). Scale-aware and definition-aware evaluation of modeled near-surface precipitation frequency using CloudSat observations. *Journal of Geophysical Research: Atmospheres*, *123*(8), 4294–4309. https://doi.org/10.1002/2017JD028213

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. https://doi.org/10.1002/jgrd.50141

Lenaerts, J. T. M., Camron, M. D., Wyburn-Powell, C. R., & Kay, J. E. (2020). Present-day and future Greenland ice sheet precipitation frequency from CloudSat observations and the community Earth system model. *The Cryosphere*, *14*(7), 2253–2265. https://doi.org/10.5194/tc-14-2253-2020

Lin, S.-J. (2004). A "Vertically Lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review*, *132*(10), 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:avlfdc>2.0.co;2

Lin, S.-J., & Rood, R. B. (1997). An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Quarterly Journal of the Royal Meteorological Society*, *123*(544), 2477–2498. https://doi.org/10.1002/qj.49712354416

Liu, X., Ma, P.-L., Wang, H., Tilmes, S., Singh, B., Easter, R. C., et al. (2016). Description and evaluation of a new four-mode version of the Modal Aerosol Module (MAM4) within version 5.3 of the Community Atmosphere Model. *Geoscientific Model Development*, *9*(2), 505–522. https://doi.org/10.5194/gmd-9-505-2016

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., et al. (2018). Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) edition-4.0 data product. *Journal of Climate*, *31*(2), 895–918. https://doi.org/10.1175/jcli-d-17-0208.1

Marchand, R. (2023). MISR Cloud-Top-Height (CTH) Optical-Depth (OD) product [Dataset]. Retrieved from https://atmos.uw.edu/~roj/MISR_observations/MISR_CTH_OD_histograms/V7

Marchand, R., Ackerman, T., Smyth, M., & Rossow, W. B. (2010). A review of cloud top height and optical depth histograms from MISR, ISCCP, and MODIS. *Journal of Geophysical Research*, *115*(D16), D16206. https://doi.org/10.1029/2009jd013422

Medeiros, B. (2023). mskd23: v1.1 [Software]. Zenodo https://doi.org/10.5281/zenodo.7971831

Medeiros, B., Kay, J. E., Shaw, J., & Davis, I. (2023). Supporting data for assessing clouds using satellite observations through three generations of global atmosphere models [Dataset]. https://doi.org/10.5281/ZENODO.7688965

Medeiros, B., Williamson, D. L., Hannay, C., & Olson, J. G. (2012). Southeast Pacific stratocumulus in the Community Atmosphere Model. *Journal of Climate*, *25*(18), 6175–6192. https://doi.org/10.1175/jcli-d-11-00503.1

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., et al. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26), eaba1981. https://doi.org/10.1126/sciadv.aba1981

MODIS Atmosphere Science Team. (2022). MODIS (Aqua/Terra) Cloud Properties Level 3 monthly, 1x1 degree grid [Dataset]. NASA Level 1 and Atmosphere Archive and Distribution System. https://doi.org/10.5067/MODIS/MCD06COSP_M3_MODIS.062

Nam, C., Bony, S., Dufresne, J. L., & Chepfer, H. (2012). The 'too few, too bright' tropical low-cloud problem in CMIP5 models. *Geophysical Research Letters*, *39*(21), L21801. https://doi.org/10.1029/2012gl053421

Noel, V., Chepfer, H., Hoareau, C., Reverdy, M., & Cesana, G. (2014). Effects of solar activity on noise in CALIOP profiles above the South Atlantic Anomaly. *Atmospheric Measurement Techniques*, *7*(6), 1597–1603. https://doi.org/10.5194/amt-7-1597-2014

Norris, J. R., & Evan, A. T. (2015). Empirical removal of artifacts from the ISCCP and PATMOS-x satellite cloud records. *Journal of Atmospheric and Oceanic Technology*, *32*(4), 691–702. https://doi.org/10.1175/jtech-d-14-00058.1

Pincus, R., Hubanks, P. A., Platnick, S. A., Meyer, K., Holz, R. E., Botambekov, D., & Wall, C. J. (2022). Updated observations of clouds by MODIS for global model assessment. *Earth System Science Data Discussions*, *2022*, 1–20.

Pincus, R., Platnick, S., Ackerman, S. A., Hemler, R. S., & Hofmann, R. J. P. (2012). Reconciling simulated and observed views of clouds: MODIS, ISCCP, and the limits of instrument simulators. *Journal of Climate*, *25*(13), 4699–4720. https://doi.org/10.1175/jcli-d-11-00267.1

Randall, D. A. (2013). Beyond deadlock. *Geophysical Research Letters*, *40*(22), 5970–5976. https://doi.org/10.1002/2013gl057998

Rasch, P. J., Xie, S., Ma, P.-L., Lin, W., Wang, H., Tang, Q., et al. (2019). An overview of the atmospheric component of the Energy Exascale Earth System Model. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2377–2411. https://doi.org/10.1029/2019ms001629

Rossow, W., Walker, A., Golea, V., Knapp, K. R., Young, A., Inamdar, A., et al. (2016). International Satellite Cloud Climatology Project Climate Data Record, H-Series [Dataset]. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5QZ281S

Rossow, W. B., Knapp, K. R., & Young, A. H. (2022). International satellite cloud climatology project: Extending the record. *Journal of Climate*, *35*(1), 141–158.

Simpson, I. R., Bacmeister, J., Neale, R. B., Hannay, C., Gettelman, A., Garcia, R. R., et al. (2020). An evaluation of the large-scale atmospheric circulation and its variability in CESM2 and other CMIP models. *Journal of Geophysical Research: Atmospheres*, *125*(13), e2020JD032835. https://doi.org/10.1029/2020jd032835

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., et al. (2013). Assessment of global cloud datasets from satellites: Project and database initiated by the GEWEX radiation panel. *Bulletin of the American Meteorological Society*, *94*(7), 1031–1049. https://doi.org/10.1175/bams-d-12-00117

Swales, D. J., Pincus, R., & Bodas-Salcedo, A. (2018). The cloud feedback model intercomparison project observational simulator package: Version 2. *Geoscientific Model Development*, *11*(1), 77–81. https://doi.org/10.5194/gmd-11-77-2018

Takahashi, H., Lebsock, M. D., Richardson, M., Marchand, R., & Kay, J. E. (2019). When will spaceborne cloud radar detect upward shifts in cloud heights? *Journal of Geophysical Research: Atmospheres*, *124*(13), 7270–7285. https://doi.org/10.1029/2018JD030242

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000jd900719

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12), eaaz9549. https://doi.org/10.1126/sciadv.aaz9549

Tselioudis, G., Rossow, W., Zhang, Y., & Konsta, D. (2013). Global weather states and their properties from passive and active satellite cloud retrievals. *Journal of Climate*, *26*(19), 7734–7746. https://doi.org/10.1175/jcli-d-13-00024.1

Tselioudis, G., Rossow, W. B., Jakob, C., Remillard, J., Tropf, D., & Zhang, Y. (2021). Evaluation of clouds, radiation, and precipitation in CMIP6 models using global weather states derived from ISCCP-H cloud property data. *Journal of Climate*, *34*(17), 7311–7324. https://doi.org/10.1175/jcli-d-21-0076.1

Ullrich, P. A., & Taylor, M. A. (2015). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part I. *Monthly Weather Review*, *143*(6), 2419–2440. https://doi.org/10.1175/mwr-d-14-00343.1

Vignesh, P. P., Jiang, J. H., Kishore, P., Su, H., Smay, T., Brighton, N., & Velicogna, I. (2020). Assessment of CMIP6 cloud fraction and comparison with satellite observations. *Earth and Space Science*, *7*(2), e2019EA000975. https://doi.org/10.1029/2019EA000975

Villani, C. (2003). *Topics in optimal transportation* (Vol. 58). American Mathematical Society.

Wu, J. (2012). *Advances in k-means clustering*. Springer-Verlag. https://doi.org/10.1007/978-3-642-29807-3

Wyant, M. C., Bretherton, C. S., Bacmeister, J. T., Kiehl, J. T., Held, I. M., Zhao, M., et al. (2006). A comparison of tropical cloud properties and responses in GCMs using mid-tropospheric vertical velocity. *Climate Dynamics*, *27*(2–3), 261–279. https://doi.org/10.1007/s00382-006-0138-4

Xie, S., Lin, W., Rasch, P. J., Ma, P.-L., Neale, R., Larson, V. E., et al. (2018). Understanding cloud and convective characteristics in version 1 of the E3SM atmosphere model. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2618–2644. https://doi.org/10.1029/2018ms001350

Yang, C. A., Diao, M., Gettelman, A., Zhang, K., Sun, J., McFarquhar, G., & Wu, W. (2021). Ice and supercooled liquid water distributions over the southern ocean based on in situ observations and climate model simulations. *Journal of Geophysical Research: Atmospheres*, *126*(24), e2021JD036045. https://doi.org/10.1029/2021jd036045

Young, A. H., Knapp, K. R., Inamdar, A., Hankins, W., & Rossow, W. B. (2018). The international satellite cloud climatology project h-series climate data record product. *Earth System Science Data*, *10*(1), 583–593. https://doi.org/10.5194/essd-10-583-2018

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., et al. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019GL085782

Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netCDF operators (NCO). *Environmental Modelling & Software*, *23*(10), 1338–1342. https://doi.org/10.1016/j.envsoft.2008.03.004

Zeng, X., & Li, X. (2023). Explicitly modeling the effects of cloud condensation nuclei on warm rain initiation. *Journal of the Atmospheric Sciences*, *80*(1), 259–272.

Zhang, Y., Xie, S., Lin, W., Klein, S. A., Zelinka, M., Ma, P.-L., et al. (2019). Evaluation of clouds in version 1 of the E3SM atmosphere model with satellite simulators. *Journal of Advances in Modeling Earth Systems*, *11*(5), 1253–1268. https://doi.org/10.1029/2018ms001562