https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020

Geoscientific Model Development
Discussions
Open Access

Evaluating Simulated Climate Patterns from the CMIP Archives Using Satellite and Reanalysis Datasets

John T. Fasullo[1]

[1]National Center for Atmospheric Research, Boulder, CO, 80302, USA

Correspondence to: John T. Fasullo (fasullo@ucar.edu)

Abstract.

An objective approach is presented for scoring coupled climate simulations through an evaluation against satellite and reanalysis datasets during the satellite era (i.e. since 1979). Here, the approach is described and applied to available Coupled Model Intercomparison Project (CMIP) archives and the Community Earth System Model Version 1 Large Ensemble archives, with the goal of benchmarking model performance and its evolution across CMIP generations. The approach adopted is designed to minimize the sensitivity of scores to internal variability, external forcings, and model tuning. Toward this end, models are scored based on pattern correlations of their simulated mean state, seasonal contrasts, and ENSO teleconnections. A broad range of feedback-relevant fields is considered and summarized on various timescales (climatology, seasonal, interannual) and physical realms (energy budget, water cycle, dynamics). Fields are also generally chosen for which observational uncertainty is small compared to model structural differences and error.

Highest mean variable scores across models are reported for well-observed fields such as sea level pressure, precipitable water, and outgoing longwave radiation while the lowest scores are reported for 500 hPa vertical velocity, net surface energy flux, and precipitation minus evaporation. The fidelity of CMIP models is found to vary widely both within and across CMIP generations. Systematic increases in model fidelity across CMIP generations are identified with the greatest improvements in dynamic and energetic fields. Examples include 500 hPa eddy geopotential height and relative humidity, and shortwave cloud forcing. Improvements for ENSO scores are substantially greater than for the annual mean or seasonal contrasts.

Analysis output data generated by this approach is made freely available online for a broad range of model ensembles, including the CMIP archives and various single-model large ensembles. These multi-model archives allow for an exploration of relationships between metrics across a range of simulations while the single-model large ensemble archives enable an estimation of the influence of internal variability on reported scores. The entire output archive, updated regularly, can be accessed at: http://webext.cgd.ucar.edu/Multi-Case/CMAT/index.html .

1 Introduction

Global climate models were first developed over half a century ago (Hunt et al. 1968, Manabe et al. 1975) and have provided insight into the climate system on a range of issues including the roles of various physical processes in the climate system and the attribution of climate events. They also are key tools for near-term initialized prediction and long-term boundary forced projections. Given their relevance for addressing issues of considerable socioeconomic importance, climate models are increasing being looked to for guiding policy-relevant decisions on long timescales and on regional levels. Many barriers exist however, chief amongst which are the biases in climate model representation of the physical system.

Adequate evaluation of climate models is nontrivial however. A key obstacle is that the longest observational records tend to monitor temperature and sea level pressure and therefore are not directly related to many of the fields thought to govern climate variability and change, such as for example cloud radiative forcing and rainfall (Burrows et al. 2018). Global direct observations of more physically relevant fields exist but are available exclusively from satellite and thus are limited in duration, with some of the most important data records beginning in recent decades. Over longer timescales, uncertainties in forcing external to the climate system further complicate model evaluation. Benchmarks of model performance must therefore be designed to deal with associated uncertainty by minimizing their influence.

1.1 Motivations

Climate modeling centers continually refine their codes with the goal of improving their models. The Climate Model Intercomparison Project (CMIP) is an effort to systematically coordinate and release targeted climate model experiments of high interest in the science community and has thus far provided three major releases, including CMIP3 (Meehl et al. 2007), CMIP5 (Taylor et al. 2012), and CMIP6 versions (Eyring et al. 2016). Major advances have also recently been made in key observationally-based climate datasets (as discussed herein). An opportunity has therefore arisen to take stock of these simulation archives and conduct a retrospective assessment of progress that has been made and challenges that remain.

While individual models are widely scrutinized, systematic surveys of model performance are relatively rare. It is the goal of this study to provide an initial benchmarking of models across CMIP generations using newly available and process-relevant observations that contextualizes model-observation differences with respect to internal variability and observational uncertainty. An additional goal is to provide related diagnostic outputs directly to the community. Both the graphical and data outputs generated may potentially be incorporated into broader community packages such as ESMValTool (Eyring et al. 2020), thus providing a unique evaluation of fully-coupled physical climate states that includes both climatological means and variability, that accounts for key uncertainties, and that benchmarks models across CMIP generations.

1.2 Challenges

A number of challenges exist for efforts aimed at comprehensively assessing climate model fidelity. Observations of many fields that are central to climate variability and change (e.g. cloud microphysics, entrainment rates, aerosol-cloud interactions) are not observed on the global, multi-decadal timescales required to comprehensively evaluate models. Fields for which observations do exist often contain observational uncertainties that are large, particularly at times when the spatial sampling of observing networks is poor (e.g. SST datasets) or for fields that contain significant uncertainty in satellite-based estimation (e.g. surface turbulent and radiative fluxes). For instances in which extended data records are unavailable, associated sensitivity to internal variability and externally imposed forcing, which also contains major uncertainties, must be considered, and evaluation of trends are particularly susceptible. In addition, model tuning approaches vary widely across centers (e.g. Schmidt et al. 2017), and in instances where climate fields are explicitly tuned, direct comparison against observations is unwarranted.

1.3 Approach

In light of these challenges and opportunities, an effort is made here to evaluate models with best-estimates of feedback relevant fields. The effort is further motivated by reported shifts in model behavior, such as for example the apparent increase in climate sensitivity to carbon dioxide in some models (Gettelman et al. 2019, Golaz et al. 2019, Neubauer et al. 2019). Do such shifts accompany systematic improvements in models and if so, in what fields? It is also of a more general interest to quantify canonical biases in models, their changes in successive model generations, and persistent biases affecting the most recent generations of climate models. The specific questions addressed here therefore include: what improvements have occurred across model generations and what persistent bias remain? What process-relevant well-observed fields are models most skillful in reproducing? To what extent are apparent improvements and persisting biases robustly detectible in the presence of internal climate variability, particularly as they relate to satellite records?

2.0 Methods

The analysis approach consists of computing a range of scores based on pattern correlations encompassing three climatic timescales: the climatological annual mean (annual), seasonal mean contrasts (JJA-DJF), and ENSO teleconnection patterns-computed from the 12-month July through June mean regressions against Niño3.4 sea surface temperatures (SST). Variables are classified according to three variable types (or realms) corresponding to the energy budget, water cycle, and dynamics. To reduce the influence of internal variability, the time period over which these fields are computed is at least 20 years, though the availability of some datasets allows for the use of longer period, further reducing susceptibility to internal variability. Contemporaneous time intervals are also selected as allowed for maximum overlap between available observations and simulated fields. The variables selected for consideration are chosen based on availability and judgement of their importance in simulating climate variability and change. In part this judgement is based on a recent community solicitation (Burrows et al. 2018) and many of the fields included are deemed by experts to be of highest relevance.

2.1 Observational Datasets

The Energy Budget Realm

Energy budget fields considered consist broadly of TOA radiative fluxes and cloud forcing, vertically integrated atmospheric energy divergence and tendency, and surface heat fluxes. Radiative fluxes at top of atmosphere (TOA) are taken from the Clouds and Earth's Radiant Energy System (CERES) Energy Balance and Filled Version 4.1 dataset (EBAFv4.1, Loeb et al. 2018). The dataset offers a number of improvements over earlier versions and datasets, with improved angular distribution models and scene identification, but is perhaps most notable for its recently updated derivation of cloud radiative forcing (CF). Historically CF has been estimated from differencing cloudy and neighboring clear regions, with the effect of aliasing meteorological contrasts between the regions (whereas models merely remove clouds from their radiative transfer scheme using collocated meteorology). In the EBAFv4.1, fields from NASA's GEOS-5 reanalysis are used to estimate fluxes and CF for collocated atmospheric conditions. From CERES, the TOA net shortwave (ASR), outgoing longwave (OLR), and net (R_T) radiative fluxes are used. In addition, estimates of shortwave CF (SW_{CF}) and longwave CF (LW_{CF}) are used.

Derived from the ERA-Interim reanalysis (Dee et al. 2011), vertical integrals of atmospheric energy are used to both assess the total energy divergence within the atmosphere ($\nabla \bullet A_E$) and its tendency ($\partial A_E / \partial t$). This provides important insight into the regional generation of atmospheric transports and their cumulative influence on the global energy budget (e.g. Fasullo and Trenberth 2008). They are also an intrinsic component necessary for computing the net surface energy fluxes (as the residual of R_T, $\nabla \bullet A_E$, and $\partial A_E / \partial t$). Given the challenges of directly observing the net surface flux, a residual method is likely the best available for large-scale evaluation of the budget. The method has been demonstrated to achieve an accuracy on par with direct observations on reginal scales and have proven superior on large scales, where the atmospheric divergences on which they rely become small, converging to zero by definition in the global mean (Trenberth and Fasullo, 2017). Uncertainty estimation of CERES fluxes is well documented (Loeb et al. 2018).

The Water Cycle Realm

Water cycle fields considered include precipitation (P), evaporation minus precipitation (EP), precipitable water (PRW), evaporation (LH), and near-surface relative humidity (RH_S). As global evaporation fields from direct observations and estimated from satellite also contain substantial uncertainty, precipitation minus evaporation is estimated instead from the vertically integrated divergence of moisture simulated in ERA-Interim, which is also arguably the most accurate means of evaluating large scale patterns and variability (Trenberth and Fasullo 2013). Precipitation is estimated from the Global Precipitation Climatology Project (Huffman et al. 2013) Climate Data Record (Adler et al. 2016). The improved version takes advantage of improvements in the gauge records used for calibration and indirect precipitation estimation from longwave radiances provided by NOAA leo-IR data. For other water cycle fields, output from the European Centre for Medium Range Weather Forecasts (ECMWF) Reanalysis Version 5 (ERA5, Hersbach et al. 2019) is used. ERA5 is the successor to ERAI, increasing the resolution of reported fields, the range of assimilated fields from recent satellite instruments, and accuracy as compared against a broad range of observations for various measures. For example, a comparison using the metrics described above applied to satellite data (CERES, GPCP) demonstrate reduced mean state annual and seasonal biases as compared to ERAI (not shown).

The Dynamical Realm

Dynamical fields considered include sea level pressure (SLP), wind speed (U_S), and 500 hPa eddy geopotential height (Z_{500}), vertical velocity (W_{500}), and relative humidity (RH_{500}). ERA5, discussed above, is used for estimation of dynamical fields, as such fields are generally not provided from satellite (excepting RH_{500}). Motivating its use, and among its notable improvements relative to earlier reanalyses, is ERA5's improved representation of the tropospheric circulation that is core to the dynamical evaluation.

2.2 Generation of Variable, Realm, and Overall Scores

Scores for annual mean, seasonal mean, and ENSO timescale metrics are generated from the area-weighted pattern correlations (Rs) between each simulated variable and the corresponding observational dataset. Weighted averages of these Rs are then used to generate a Variable Score and for each simulation. Averages across the relevant Variable Scores are then used to generate Realm Scores, and the Realm Scores are averaged to generate an Overall Score. Timescale Scores are also generated by averaging Rs across variables for each timescale metric. The inclusion of both Realm and Timescale scores is motivated in part by the need to interpret the origin of changes in Overall Scores, which include a large number of Rs that may otherwise obscure an obvious physical interpretation for the Overall Score.

The use of weights in generating Variable Scores is motivated by the desire to promote interpretation of differences in the Overall Score relative to the influence of internal variability. Using the Community Earth System Version 1 Large Ensemble (CESM1-LE, Kay et al. 2015), weights for ENSO scores are reduced to 0.978 (while for annual and seasonal means they are 1.00) such that the standard deviation range in Overall Scores for the 40 members of the CESM1-LE is 0.01. This therefore can be used to interpret generally the approximate contribution of internal variability to inter-model Overall Scores in analysis of the CMIP archives, suggesting that differences between individual simulations of less than approximately 0.04 are insignificant. Where available, multiple-simulation analyses provide an opportunity for further narrowing the accuracy of statements regarding inter-model comparisons of fidelity that can be made, and as will be seen, Overall Score ranges within and across the CMIP ensembles generally exceed the obscuring effects of internal variability.

2.3 CMIP Simulations

As the goal of this work is to characterize the evolution of agreement between climate models generally across the CMIP archives, and observations, all available model submissions for which sufficient data fields are provided are included in the analysis (as summarized in Table 1). A major exception to the data availability requirement relates to near surface wind speed (U_S), which was not included as part of the CMIP3 variable list specification. Scores for the dynamical realm in CMIP3 therefore omit U_S as a scored variable and instead compute the dynamic Realm score from the remaining dynamic variable scores. While multiple ensemble members are provided in the CMIP archives for many models, and have been assessed, only a single member of each model is incorporated into the analysis here to avoid overweighting the influence of any single mode.

3.0 Assessing CMIP Scores

To illustrate the analysis approach and provide context for the magnitude of biases relative to internal variability and observational uncertainty, Figure 1 shows both observed and simulated SW_{CF} fields across the timescales considered (Fig. 1a, annual, 1b) seasonal, and 1c) ENSO) in the CESM Version 2 submission to CMIP6, CERES estimates (Fig. 1d-f), and their differences (Fig. 1g-i). Significant spatial structure characterizes all fields, with a strong SW_{CF} cooling influence in the mean across much of the globe (Fig. 1a), seasonal contrasts (Fig. 1b) that vary between land and ocean and latitudinal zone, and ENSO teleconnections (Fig. 1c) that extend from the tropical Pacific Ocean to remote ocean basins and the extratropics. While (as will be seen), CESM2 scores among the best available climate models in CMIP6, large model-observation differences nonetheless exist. Regions where model-observation differences are larger than internal variability in the annual and seasonal means (stippled) are widespread and remain extensive where the uncertainty range is expanded to incorporate estimated observational uncertainty (hatched) from Loeb et al. 2018. Of particular note is the fact that it is the large-scale patterns of robust bias, where model-observational disagreement exceeds uncertainty bounds, that are the primary drivers of pattern correlations used in scoring. These are then combined into various aggregate measures, which include Variable, Realm, and Overall Scores.

The color table summary of scores for CMIP3 (Figure 2) provides a visual summary of simulation performance across the models in the archive (abscissa), including Variable, Realm, Timescale, and Overall Scores (i.e. aggregate scores, ordinate). Simulations are sorted by Overall Scores (top row, descending scores toward right). Realm and Timescale Scores (rows 2 through 7) also provide broad summaries of model performance. Mean Overall Scores (69±7, 1 sigma) are modest generally in CMIP3 and generally uniform across realms. CMIP3 simulations score particularly poorly for ENSO, where scores average to 47, are generally less than 60, and approach 0 in some models. Variable scores are highest for SLP, and PRW and OLR which are strongly tied to surface temperature, and less for other variables, with the lowest scores reported for R_S and W500. Spread across models for R_S is particularly large relative to other variables. Average scores are also poor for SW_{CF}

(68), LW_{CF} (71), and P (69), which are among the more important simulated fields according to expert consensus (Burrows et al. 2018).

The color table summary of scores for CMIP5 (Figure 3) reveals scores that are considerably higher than most CMIP3 simulations, with improvements in the average Overall Score of (75±5) and most notable improvements on the ENSO timescale, with an average of 57, though with considerable inter-model range (σ=10). A broad increase in scores in the highest performing models is apparent with numerous variable scores exceeding 85 (orange/red). As for CMIP3 the highest scoring variables are PRW, SLP, and OLR with RHS and W500 being the lowest scoring variables. Scores remain relatively low for SW_{CF} (71), LW_{CF} (75), and P (73).

The color table summary of scores for CMIP6 (Figure 4) illustrates scores that are considerably higher than both CMIP3 and CMIP5 simulations, with improvements in the average Overall Score of (79±4) and most continued improvements on the ENSO timescale, though again with considerable inter-model range. A continued increase in scores in the highest performing models is again apparent, with scores reaching the mid- to upper 70s and numerous variable scores exceeding 90 (red). The highest scoring variables again include PRW, SLP, and OLR though scores are also high for RH_{500}, one of the more important simulated fields according to expert consensus (Burrows et al., 2018). Scores also increase for SW_{CF} (78), LW_{CF} (80), and P (77).

To highlight connections between variables, and the main variables driving variance in aggregate scores across the CMIP archives, cross correlations are shown in Figure 5. Correlations between variables and realms reveal variables that exhibit strong connections to other variables and aggregate scores. For Overall Scores, these include strong connections to P, E-P and OLR, fields strongly connected to atmospheric heating, dynamics, and deep convection and therefore broadly relevant to all realms considered. Strong connections also exist for SW_{CF}, LW_{CF}, and RH_{500}, consistent with the expert consensus in highlighting these fields are being particularly import (Burrows et al. 2018). An approximately equal correlation exists across Realms with the Overall Score, while for timescales, ENSO exhibits the strongest overall correlation as it contains the greatest inter-model variance and thus explains a greater portion of the Overall Score variance. Notable as well is that some variables for which scores are high in the mean, such as SLP and PRW, exhibit little correlation with the Overall Score as the uniformly high scores across models impart relatively little variance to the spread in Overall Scores across models.

4.0 Derived Bias Patterns for Select Variables

The observational estimate for SW_{CF} from CERES is shown in Figure 6a along with mean bias patterns for CMIP3 (b) and CMIP6 (c). A principal component (PC) analysis of the bias across the CMIP archives is also shown with the leading

https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020

principal components and their tercile mean values within each CMIP version being shown (d) along with the characteristics of the two leading patterns of bias (Fig. 6d-f). In the PC analysis, the observational benchmark field is also included to gauge improvements or degradation of model PCs across CMIP generations. The mean observational field (Fig. 6a) is characterized by negative values in nearly all locations (except over ice) and the strongest cooling influence in the deep tropics, subtropical stratocumulus regions, and midlatitude oceans. Mean bias patterns demonstrate considerable improvement across the CMIP generations, with major reductions in negative biases in the subtropics and tropics over ocean. Variance across models is characterized by differing tropical-extratropical contrasts in SW_{CF} (EOF1), which explain 24% of the inter-model variance, and land-ocean contrasts (EOF2), which explain 16% of the variance. The expression of both patterns of biases is demonstrated to diminish across CMIP generations and terciles in their PC weights (Fig. 6d), with CMIP6 values lying closer to observational estimates than CMIP PC1/2 weights. Improvements are not in general monotonic across the CMIP generations, with improvements and degradations notable in some aspects of the PC1/2 transition from CMIP3 to CMIP5.

The observational estimate for LW_{CF} from CERES is shown in Figure 7a along with mean bias patterns for CMIP3 (b) and CMIP6 (c). A PC analysis of the bias across the CMIP archives is also shown with the leading PC weights and their tercile mean values within each CMIP version being shown (d) along with the two leading patterns of bias (Fig. 7e, f). Observational fields are characterized by a strong heating influence in regions of deep tropical convection and in the extratropical ocean regions in which SW_{CF} was also strong while weak heating is evident in the subtropics and polar regions. Significant changes characterize mean bias patterns between CMIP3 and CMIP6, with positive biases across most ocean regions in CMIP3 and negative biases in many of the same regions in CMIP6. On average however, the magnitude of biases are reduced across CMIP generations. This is evident for example in the PC analysis of bias (Fig. 7e-f), where the leading mode (EOF1, Fig. 7e) exhibits strong weightings over the warm pool, is negatively correlated with both the mean pattern and bias, and explains 37% of the inter-model variance. In contrast, EOF2 exhibits a strong tropical-extratropical contrast and explains only 13% of the bias variance. The PC1/2 tercile weights for these modes show a considerable reduction in EOF1 spread and lower tercile bias and generally improved agreement across model terciles from CMIP3 to CMIP6, though as with SW_{CF}, the improvement is not monotonic nor uniform across all terciles and PCs.

The observational estimate for precipitation from GPCP is shown in Figure 8a along with mean bias patterns for CMIP3 (b) and CMIP6 (c). The PC analysis of the bias across the CMIP archives is also shown with the leading PC tercile mean values for each CMIP version being shown (d) along with the two leading patterns of bias (Fig. 8e, f). The annual mean pattern resolves key features of the climate system, including strong precipitation in the Inter-Tropical Convergence Zone (ITCZ) and low precipitation in the subtropics and at high latitudes. Biases are large in both CMIP3 and CMIP6 on average and are characterized generally by excessive subtropical precipitation and deficient precipitation in the ITCZ, South America, and at high latitudes. Earlier work has generally characterized model bias in terms of its ITCZ structure (Oueslati et al. 2015), an important aspect of the bias, though systematic bias is also apparent here outside of the tropical Pacific. In addition, the PC decomposition of CMIP precipitation biases (Fig. 8d-f) suggests that the bias is comprised to two distinct leading patterns that together explain 15% and 11% of the variance across models (i.e. a separable single leading pattern is not starkly evident). The first pattern is characterized by weakness in precipitation across the equatorial oceans, with elevated rates in the Maritime continent and in the Pacific Ocean near 15N/S. The second mode of precipitation bias is characterized by loadings over Africa and South America, and just south of the climatological Pacific ITCZ location (Fig. 8a), with negative values in the subtropical ocean basins. Based on the evolution of weights in PC terciles, slight improvement across CMIP generations is evident, as tercile values lie closer to observations for all terciles of PC1/2 in CMIP6 versus CMIP3, with the exception of the upper terciles of PC2 and the lower terciles of PC1.

The observational estimate for RH_{500} from ERA5 is shown in Figure 9a along with mean bias patterns for CMIP3 (b) and CMIP6 (c). A principal component analysis of the bias across the CMIP archives is also shown with the leading principal components and their tercile mean values within each CMIP version being shown (d) along with the two leading patterns of bias (Fig. 9e, f). The observed RH_{500} field is characterized by very dry conditions in the subtropics, with values generally below 30% across broad regions that were largely unresolved in CMIP3 (e.g. Fasullo and Trenberth 2012), and positive humidity biases in regions of frequency deep convection (i.e. Maritime Continent, Amazon) and at high latitudes. The CMIP3 mean bias field is negatively correlated with the mean state, with patterns that lack sufficient spatial variability, are too moist in the subtropics, and too dry in the Maritime continent, the Amazon, and at high latitudes. The magnitude of mean RH_{500} biases in CMIP6 are substantially smaller (roughly 50%) than CMIP3, though a similar overall pattern exists. The PC analysis of bias reveals a leading pattern of bias that explains 50% of the intermodal variance and is positively correlated with both the CMIP-mean bias and observed mean field (0.45). The second leading pattern (Fig. 9f) explains considerably less variance (14%) and exhibits a zonally uniform structure characterized by tropical-extratropical contrast. The weights for PC1/2 reveal systematic bias in PC1 across models, and considerable improvement across CMIP generations as CMIP6 weights lie significantly closer to observations that CMIP3 weights for all terciles. Very slight corresponding improvement in PC2, while suggested for the upper terciles, is not however evident in the lower tercile of models, though this comprises a small fraction of variance in CMIP bias.

In the effort to summarize the evolution of the full distributions of scores across the CMIP archives, whisker plots encompassing the median, interquartile, and 10th-90th percentile ranges are shown for various aggregate metrics and key fields in Figure 10. Also shown are the equivalent ranges for scores computed from the CESM1-LE to provide context for the uncertainty in scores associated with internal variability for each distribution. A steady progression in the Overall Scores is evident across CMIP versions. The improvements are also evident across Realm Scores and particularly for the poorest scoring models in the Dynamics Realm. Scores for Annual and Seasonal timescales are generally high across archives, though internal variability is also small and is substantially less than the median improvements across archives. The range of scores for ENSO is significantly greater than other timescales, as is the range of internal variability, and substantial improvements have been realized for the lowest scoring models across successive CMIP generations. Noteworthy are the substantial improvements in SW_{CF}, LW_{CF}, and P, with the best CMIP3 simulations scoring near the median value for CMIP6 and changes in median values exceeding uncertainty arising from internal variability. Scores for RH_{500} have also improved, although the spread within the CMIP3 archives is substantial and uncertainty arising from internal variability is somewhat greater than for other variables, and RH_{500} scores are generally higher than for cloud forcing and P. For SLP, median scores are uniformly high across the CMIP generations, with small but steady improvement in median and interquartile scores, with the main exception being the low scoring 10-25% range of CMIP3 simulations.

5.0 Discussion

An objective model evaluation tool has been developed that uses feedback-relevant fields and takes advantage of recent advances in satellite and reanalysis observations. In its application to the CMIP archives, the tool is shown to be useful for computing model scores across variables, realms, and timescales, using the best available satellite and observational estimates of present-day climate. The tool also provides visual summaries of model performance across the CMIP archives, which readily allow for the survey of a broad suite of climate performance scores.

Based on the pattern correlation approach adopted, a number of statements can be made regarding the overall performance of climate models across CMIP generations. Also noteworthy is that, as gauged by analysis of the CESM1-LE, and consistent with the motivations and design of the approach used here, these statements are robust to the obscuring influence of internal climate variability. In general, computed scores have increased steadily across CMIP generations, with improvements exceeding the uncertainty associated with internal variability. Associated with these improvements, the leading patterns of bias across models are shown to have been reduced. Improvements are large and particularly noteworthy for ENSO teleconnection patterns, as the poorest scoring models in each CMIP generation have improved substantially. The overall range of model performance has also decreased in conjunction with increases in median scores, as improvements in the worst models has outpaced that of the median. Reductions in systematic patterns of bias (e.g. Figs. 6-9) across the CMIP archives have been particularly pronounced for fields deemed in expert solicitations to have disproportionate importance, including SW_{CF}, LW_{CF}, and RH_{500}.

Also relevant for climate feedbacks, Variable Scores for SW_{CF}, LW_{CF}, RH_{500}, and precipitation have increased steadily across the CMIP generations (e.g. Fig. 10), with magnitudes exceeding the uncertainty associated with internal variability. Scores are particularly high for CMIP6 models for which high climate sensitivities have been reported, including CESM2, SAM0-UNICON, GFDL-CM4, CNRM-CM6-1, E3SM, and EC-Earth3-Veg (though exceptions also exist such as in the case of MIRCO6). These findings therefore echo the concerns voiced in Gettelman et al. 2019: "What scares us is not that the CESM2 ECS is wrong (all models are wrong, (Box, 1976)) but that it might be right.". Further work examining the ties between metrics of performance in simulating the present-day climate, such as those provided here, and longer-term climate model behavior is warranted to bolster confidence in model projections of climate change.

Data Availability

Data used in this study are available freely from the Earth System Grid at: https://www.earthsystemgrid.org

NetCDF output for the fields generated herein is freely available at: http://webext.cgd.ucar.edu/Multi-Case/CMAT/index.html

Acknowledgements

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. Portions of this study were supported by the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program of the U.S. Department of Energy's Office of Biological & Environmental Research (BER) via National Science Foundation IA 1844590.

References

Adler, R., Sapiano, M., Huffman, G., Bolvin, D., Gu, G., Wang, J., ... and Schneider, U.: The new version 2.3 of the Global Precipitation Climatology Project (GPCP) monthly analysis product. University of Maryland, April, 1072-1084, 2016.

Box, G. E. P.: Science and statistics. J. Amer. Statistical Assoc., 71(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949, 1976.

Burrows, S. M., Dasgupta, A., Reehl, S., Bramer, L., Ma, P. L., Rasch, P. J., and Qian, Y.: Characterizing the relative importance assigned to physical variables by climate scientists when assessing atmospheric climate model fidelity. Adv. Atm. Sci., 35(9), 1101-1113, doi: doi:10.1007/s00376-018-7300-x, 2018.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... and Bechtold, P.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quart. J. Roy Met. Soc., 137(656), 553-597, doi: 10.1002/qj.828, 2011.

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

Eyring, V. and Coauthors: ESMValTool v2.0 – Extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geosci. Model Dev. Disc., in review, 2020.

Fasullo, J. T. and Trenberth, K. E. (2008). The annual cycle of the energy budget. Part I: Global mean and land–ocean exchanges. J. Clim., 21(10), 2297-2312, doi: 10.1175/2007JCLI1935.1, 2008.

Fasullo, J. T., and Trenberth, K. E.: A less cloudy future: The role of subtropical subsidence in climate sensitivity. Science, 338(6108), 792-794, doi: 10.1126/science.1227465, 2012.

Gettelman, A., Hannay, C., Bacmeister, J.T., Neale, R., Pendergrass, A.G., Danabasoglu, G., ... Mills, M.J.: High climate sensitivity in the Community Earth System Model Version 2 (CESM2). Geophys. Res. Lett., 46(14), 8329-8337 doi: 10.1029019GL083978, 2019.

Golaz, J.C, et al.: The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution." J. of Adv. in Modeling Earth Systems 11.7, 2089-2129, doi: 10.1029/2018MS001603, 2019.

Hersbach, H., and Coauthors: Global reanalysis: goodbye ERA-Interim, hello ERA5. ECMWF, doi:10.21957/vf291hehd7. https://www.ecmwf.int/node/19027, 2019.

Huffman, G. J., Adler, R.F., Bolvin, D.T. and Gu G.: Improving the global precipitation record: GPCP version 2.1. Geophys., Res., Lett., 36, L17808, doi:10.1029/2009GL040000, 2009.

Hunt, B. G., and Manabe S.: Experiments with a stratospheric general circulation model: II. Large-scale diffusion of tracers in the stratosphere. Monthly Weather Review 96.8 (1968): 503-539, doi: 10.1175/1520-0493(1968)096<0503:EWASGC>2.0.CO;2, 2009.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand G., ... and Holland, M.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. Bull. Amer. Met. Soc., 96(8), 1333-1349, doi:10.1175/BAMS-D-13-00255.1, 2015.

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., ... and Kato, S.: Clouds and the earth's radiant energy system (CERES) energy balanced and filled (EBAF) top-of-atmosphere (TOA) edition-4.0 data product. J. Clim., 31(2), 895-918, doi: 10.1175/JCLI-D-17-0208.1, 2018.

Manabe, S., Bryan, K., and Spelman, M. J.: A global ocean-atmosphere climate model. Part I. The atmospheric circulation. J. Phys. Ocn., 5(1), 3-29, doi: 10.1175/1520-0485(1975)005<0003:AGOACM>2.0.CO;2, 1975.

Meehl, G. A. and Coauthors: The WCRP CMIP3 multimodel dataset: a new era in climate change Research. Bull. Am. Met.. Soc. 88, 1383–1394 (2007), doi: 10.1175/JCLI3675.1, 2007.

Neubauer, D., Ferrachat, S., Drian, S. L., Stier, P., Partridge, D. G., Tegen, I., ... and Lohmann, U.: The global aerosol-climate model ECHAM6. 3-HAM2. 3–Part 2: Cloud evaluation, aerosol radiative forcing and climate sensitivity. Geosci. Mod. Dev. Disc., doi: 10.5194/gmd-12-3609-2019, 2019.

Oueslati, B., and Bellon, G.: The double ITCZ bias in CMIP5 models: interaction between SST, large-scale circulation and precipitation. Clim. Dyn., 44(3-4), 585-607, doi: 10.1007/s00382-015-2468-6, 2015.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J. C., Hannay, C., ... and Saha, S.: Practice and philosophy of climate model tuning across six US modeling centers. Geosci. Model Dev. Disc., 10(9), 3207, doi: 10.5194/gmd-10-3207-2017, 2017.

Taylor, K. E., R.J. Stouffer, and Meehl G.A.: An overview of CMIP5 and the experiment design, Bull. Amer. Met. Soc. 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Trenberth, K. E., and Fasullo, J. T.: Regional energy and water cycles: Transports from ocean to land. J. Clim., 26(20), 7837-7851, doi: 10.1175/JCLI-D-13-00008.1, 2013.

Trenberth, K. E., and Fasullo, J. T.: Atlantic meridional heat transports computed from balancing Earth's energy locally. Geo. Res. Lett. 44(4), 1919-1927, doi: 10.1002/2016GL072475, 2017.

Tables

Table 1: Sorted summary of CMIP models considered in this work, sorted by Overall Scores.

https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020

Geoscientific Model Development Discussions
Open Access

CMIP3	CMIP5	CMIP6
gfdl_cm2_0 (0.78)	CESM1-BGC (0.81)	CESM2 (0.86)
gfdl_cm2_1 (0.75)	CNRM-CM5-2 (0.81)	MIROC6 (0.85)
cccma_cgcm3_1_t63 (0.75)	CESM1-FASTCHEM (0.81)	CESM2-WACCM (0.85)
mri_cgcm2_3_2a (0.75)	CESM1-CAM5 (0.81)	GISS-E2-1-H (0.85)
mpi_echam5 (0.75)	ACCESS1-0 (0.81)	SAM0-UNICON (0.84)
miub_echo_g (0.74)	NorESM1-ME (0.80)	GFDL-CM4 (0.84)
csiro_mk3_5 (0.74)	CESM1-WACCM (0.80)	EC-Earth3-Veg (0.84)
ingv_echam4 (0.73)	CESM1-CAM5-1-FV2 (0.80)	EC-Earth3 (0.83)
ukmo_hadcm3 (0.73)	MIROC5 (0.80)	UKESM1-0-LL (0.82)
cccma_cgcm3_1 (0.73)	CMCC-CMS (0.80)	MRI-ESM2-0 (0.82)
cnrm_cm3 (0.73)	HadGEM2-ES (0.80)	E3SM-1-0 (0.81)
ncar_ccsm3_0 (0.72)	NorESM1-M (0.79)	CNRM-CM6-1 (0.81)
csiro_mk3_0 (0.71)	BNU-ESM (0.79)	CNRM-ESM2-1 (0.81)
miroc3_2_medres (0.71)	ACCESS1-3 (0.78)	MIROC-ES2L (0.81)
bccr_bcm2_0 (0.71)	HadGEM2-AO (0.78)	FGOALS-g3 (0.79)
iap_fgoals1_0_g (0.69)	bcc-csm1-1-m (0.77)	CAMS-CSM1-0 (0.79)
miroc3_2_hires (0.69)	GFDL-CM2p1 (0.76)	BCC-CSM2-MR (0.77)
ukmo_hadgem1 (0.68)	CanESM2 (0.76)	BCC-ESM1 (0.77)
ipsl_cm4 (0.67)	CMCC-CESM (0.75)	CanESM5 (0.77)
ncar_pcm1 (0.61)	IPSL-CM5B-LR (0.75)	IPSL-CM6A-LR (0.74)
inmcm3_0 (0.60)	MRI-ESM1 (0.75)	GISS-E2-1-G (0.74)
giss_model_e_r (0.60)	MPI-ESM-LR (0.75)	NorESM2-LM (0.74)
giss_aom (0.59)	MPI-ESM-MR (0.74)	
giss_model_e_h (0.46)	MPI-ESM-P (0.74)	
	MRI-CGCM3 (0.74)	
	FGOALS-g2 (0.74)	
	GFDL-ESM2G (0.72)	
	GISS-E2-R-CC (0.72)	
	IPSL-CM5A-MR (0.71)	
	MIROC-ESM (0.70)	
	GISS-E2-H-CC (0.69)	
	IPSL-CM5A-LR (0.68)	
	CSIRO-Mk3-6-0 (0.68)	
	MIROC-ESM-CHEM (0.68)	
	inmcm4 (0.68)	
	GISS-E2-H (0.67)	
	CESM1-BGC (0.81)	
	CNRM-CM5-2 (0.81)	
	CESM1-FASTCHEM (0.81)	
	CESM1-CAM5 (0.81)	
	ACCESS1-0 (0.81)	
	NorESM1-ME (0.80)	
	CESM1-WACCM (0.80)	
	CESM1-CAM5-1-FV2 (0.80)	
	MIROC5 (0.80)	
	CMCC-CMS (0.80)	
	HadGEM2-ES (0.80)	
	NorESM1-M (0.79)	
	BNU-ESM (0.79)	
	ACCESS1-3 (0.78)	
	HadGEM2-AO (0.78)	
	bcc-csm1-1-m (0.77)	
	GFDL-CM2p1 (0.76)	
	CanESM2 (0.76)	
	CMCC-CESM (0.75)	
	IPSL-CM5B-LR (0.75)	
	MRI-ESM1 (0.75)	
	MPI-ESM-LR (0.75)	
	MPI-ESM-MR (0.74)	
	MPI-ESM-P (0.74)	
	MRI-CGCM3 (0.74)	
	FGOALS-g2 (0.74)	
	GFDL-ESM2G (0.72)	
	GISS-E2-R-CC (0.72)	
	IPSL-CM5A-MR (0.71)	
	MIROC-ESM (0.70)	
	GISS-E2-H-CC (0.69)	
	IPSL-CM5A-LR (0.68)	
	CSIRO-Mk3-6-0 (0.68)	
	MIROC-ESM-CHEM (0.68)	
	inmcm4 (0.68)	
	GISS-E2-H (0.67)	

https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020

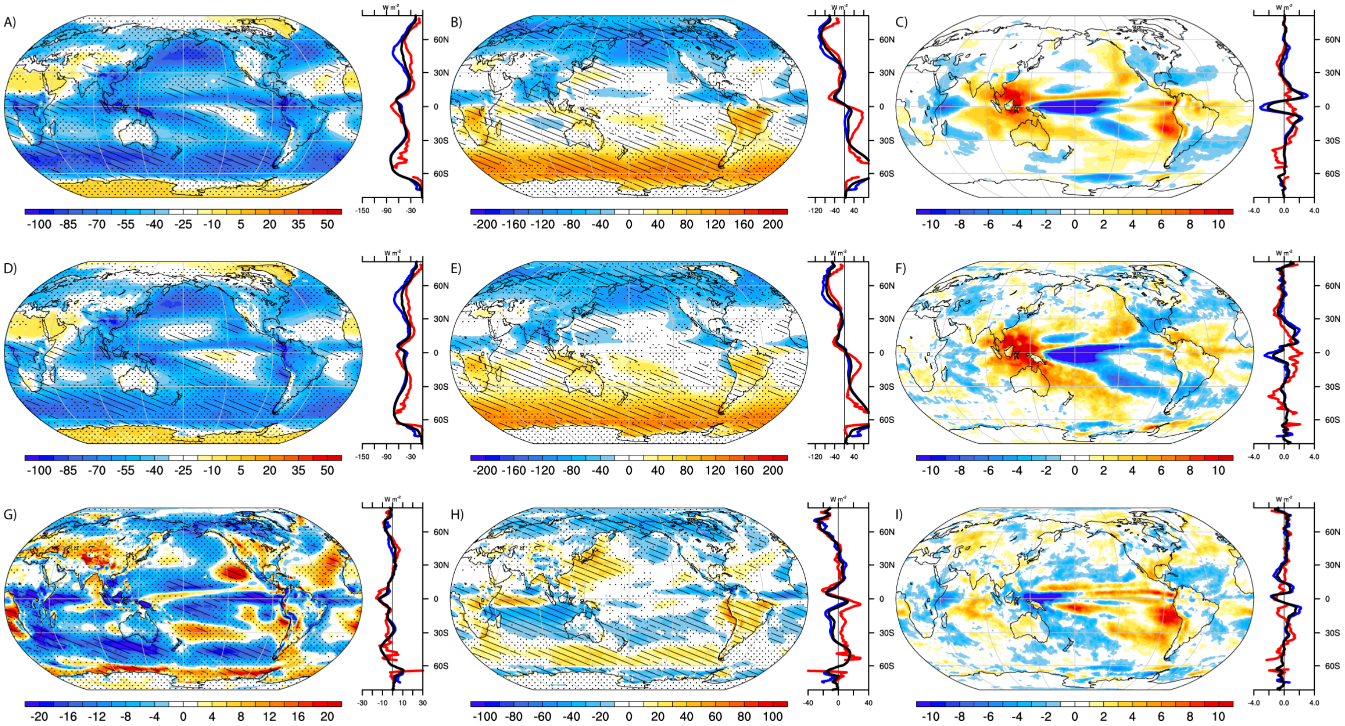
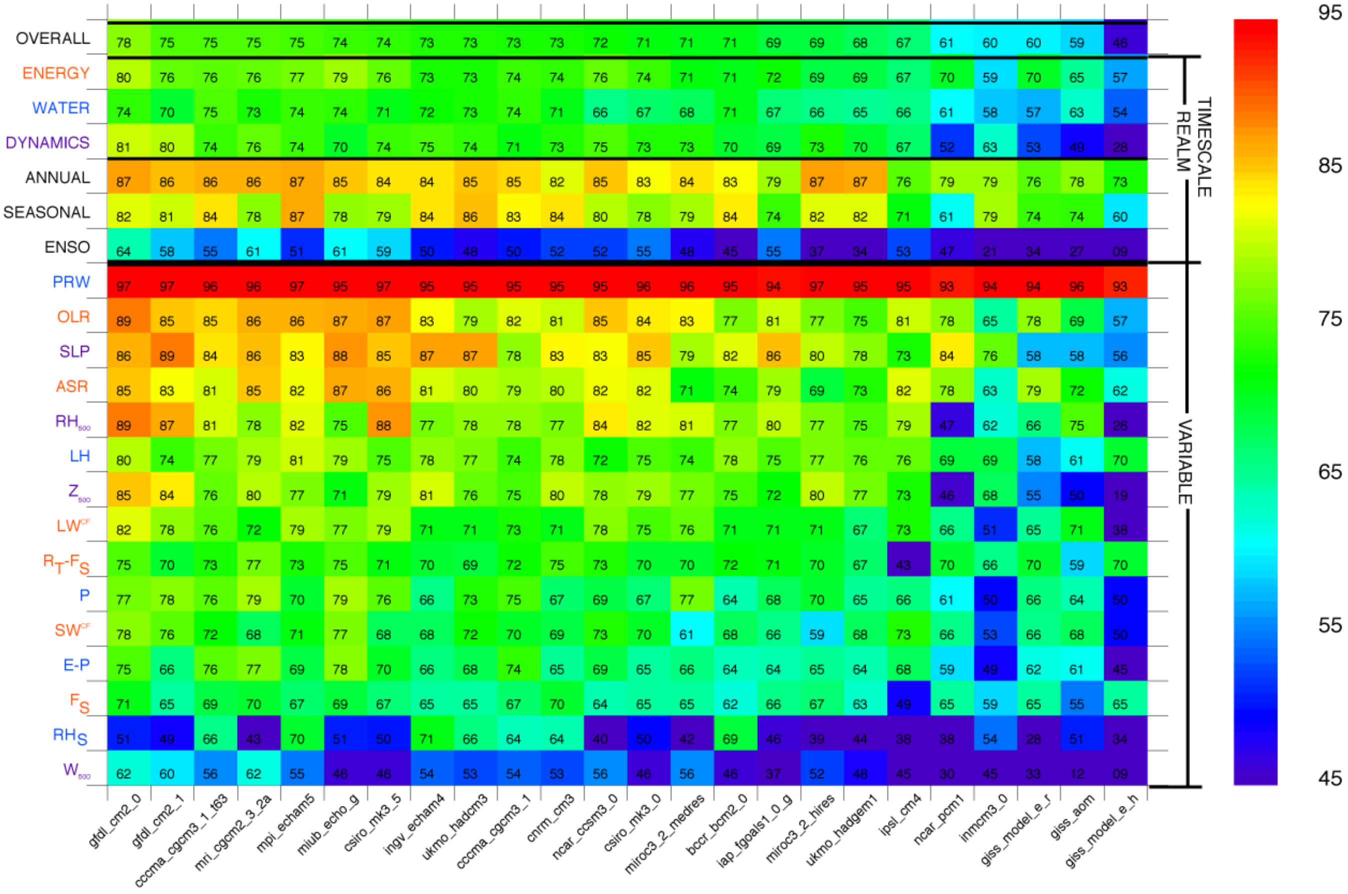


Figure 1: Mean simulated fields of SW_{CF} in CESM2 from 1995-2014 for A) the annual mean, B) seasonal contrasts, and C) regressed against Niño3.4 SST anomalies from July-June. Observed CERES EBAF4.1 estimated SW_{CF} for 2000-2018 for analogous metrics (D-F) where stippling indicates regions where CESM-CERES differences exceed twice the estimated internal spread from CESM-LE. Hatching indicates regions where these differences (G-I) exceed the same spread and observational uncertainty (added in quadrature, applied to all panels in each column). Units are W m^{-2} except for right column where units are W m^{-2} K^{-1}.

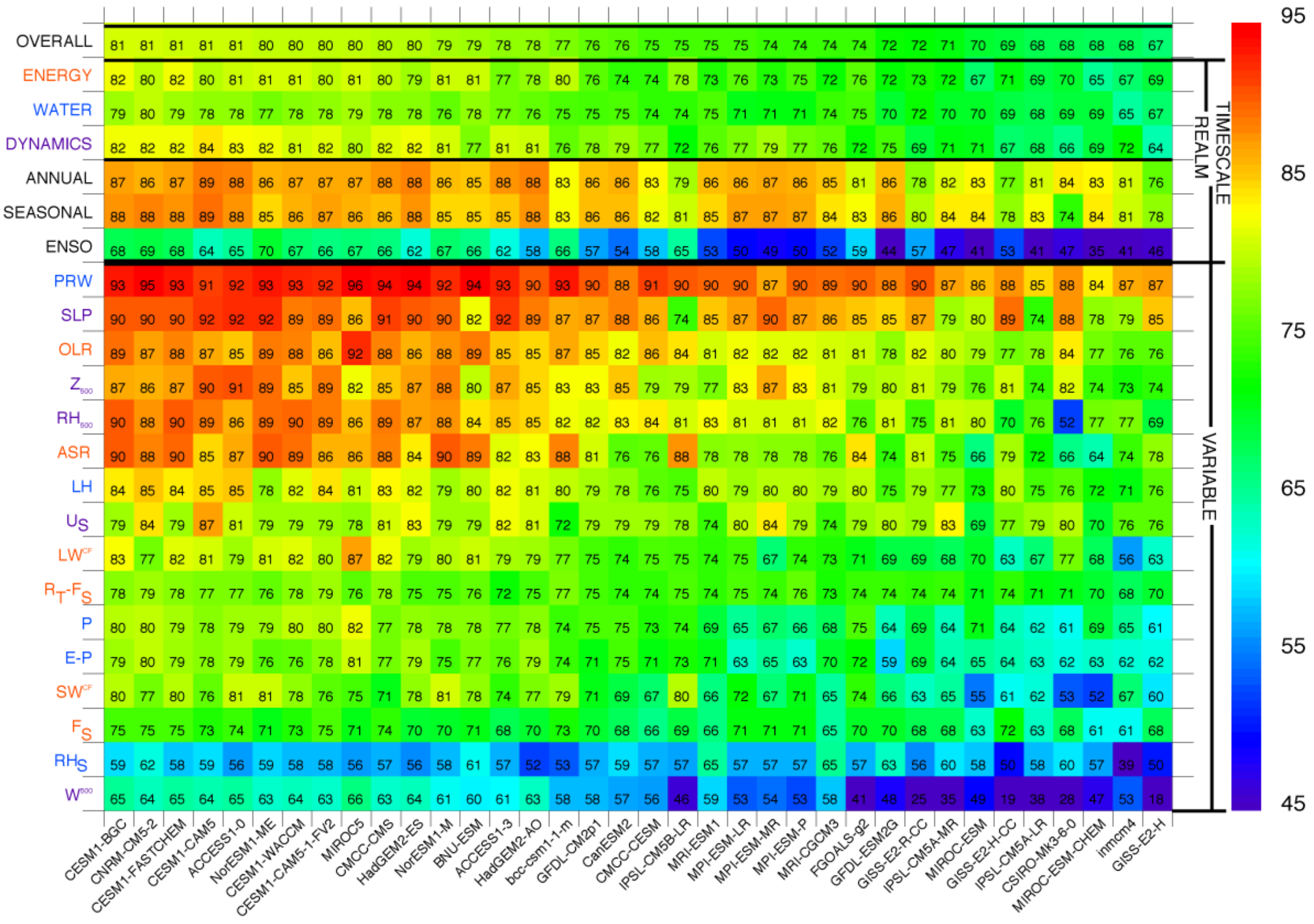
https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020



	gfdl_cm2_0	gfdl_cm2_1	cccma_cgcm3_1_t63	mri_cgcm2_3_2a	mpi_echam5	miub_echo_g	csiro_mk3_5	ingv_echam4	ukmo_hadcm3	cccma_cgcm3_1	cnrm_cm3	ncar_ccsm3_0	csiro_mk3_0	miroc3_2_medres	bccr_bcm2_0	iap_fgoals1_0_g	miroc3_2_hires	ukmo_hadgem1	ipsl_cm4	ncar_pcm1	inmcm3_0	giss_model_e_r	giss_aom	giss_model_e_h
OVERALL	78	75	75	75	75	74	74	73	73	73	73	72	71	71	71	69	69	68	67	61	60	60	59	46
ENERGY	80	76	76	76	77	79	76	73	73	74	74	76	74	71	71	72	69	69	67	70	59	70	65	57
WATER	74	70	75	73	74	74	71	72	73	74	71	66	67	68	71	67	66	65	66	61	58	57	63	54
DYNAMICS	81	80	74	76	74	70	74	75	74	71	73	75	73	73	70	69	73	70	67	52	63	53	49	28
ANNUAL	87	86	86	86	87	85	84	84	85	85	82	85	83	84	83	79	87	87	76	79	79	76	78	73
SEASONAL	82	81	84	78	87	78	79	84	86	83	84	80	78	79	84	74	82	82	71	61	79	74	74	60
ENSO	64	58	55	61	51	61	59	50	48	50	52	52	55	48	45	55	37	34	53	47	21	34	27	09
PRW	97	97	96	96	97	95	97	95	95	95	95	95	96	96	95	94	97	95	95	93	94	94	96	93
OLR	89	85	85	86	86	87	87	83	79	82	81	85	84	83	77	81	77	75	81	78	65	78	69	57
SLP	86	89	84	86	83	88	85	87	87	78	83	83	85	79	82	86	80	78	73	84	76	58	58	56
ASR	85	83	81	85	82	87	86	81	80	79	80	82	82	71	74	79	69	73	82	78	63	79	72	62
RH_{500}	89	87	81	78	82	75	88	77	78	78	77	84	82	81	77	80	77	75	79	47	62	66	75	26
LH	80	74	77	79	81	79	75	78	77	74	78	72	75	74	78	75	77	76	76	69	69	58	61	70
Z_{500}	85	84	76	80	77	71	79	81	76	75	80	78	79	77	75	72	80	77	73	46	68	55	50	19
LW^{CF}	82	78	76	72	79	77	79	71	71	73	71	78	75	76	71	71	71	67	73	66	51	65	71	38
R_T-F_S	75	70	73	77	73	75	71	70	69	72	75	73	70	70	72	71	70	67	43	70	66	70	59	70
P	77	78	76	79	70	79	76	66	73	75	67	69	67	77	64	68	70	65	66	61	50	66	64	50
SW^{CF}	78	76	72	68	71	77	68	68	72	70	69	73	70	61	68	66	59	68	73	66	53	66	68	50
E-P	75	66	76	77	69	78	70	66	68	74	65	69	65	66	64	64	65	64	68	59	49	62	61	45
F_S	71	65	69	70	67	69	67	65	65	67	70	64	65	65	62	66	67	63	49	65	59	65	55	65
RH_S	51	49	66	43	70	51	50	71	66	64	64	40	50	42	69	46	39	44	38	38	54	28	51	34
W_{500}	62	60	56	62	55	46	46	54	53	54	53	56	46	56	46	37	52	48	45	30	45	33	12	09

Figure 2: Overall, Realm, Timescale, and Variable scores (ordinate) for historical (20c3m) simulations submitted to the CMIP3 archives (abscissa) sorted by overall score (top row) based on methods employed (see text). Simulations and variables are ordered in descending score order from left to right using overall score and from top to bottom using average variable score, respectively.

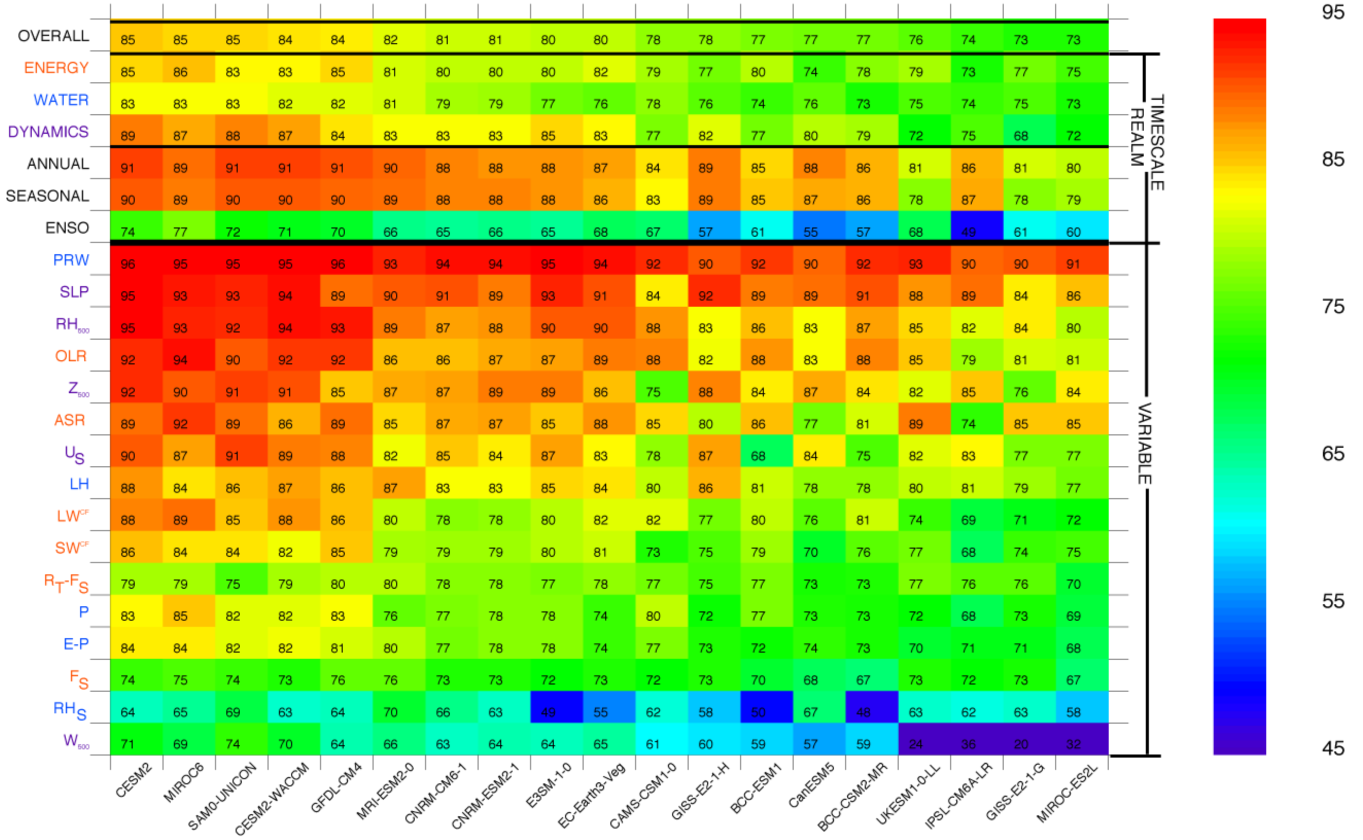
https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020



	CESM1-BGC	CNRM-CM5-2	CESM1-FASTCHEM	CESM1-CAM5	ACCESS1-0	NorESM1-ME	CESM1-WACCM	CESM1-CAM5-1-FV2	MIROC5	CMCC-CMS	HadGEM2-ES	NorESM1-M	BNU-ESM	ACCESS1-3	HadGEM2-AO	bcc-csm1-1-m	GFDL-CM2p1	CanESM2	CMCC-CESM	IPSL-CM5B-LR	MRI-ESM1	MPI-ESM-LR	MPI-ESM-MR	MPI-ESM-P	MRI-CGCM3	FGOALS-g2	GFDL-ESM2G	GISS-E2-R-CC	IPSL-CM5A-MR	MIROC-ESM	GISS-E2-H-CC	IPSL-CM5A-LR	CSIRO-Mk3-6-0	MIROC-ESM-CHEM	inmcm4	GISS-E2-H
OVERALL	81	81	81	81	81	80	80	80	80	80	80	79	79	78	78	77	76	76	75	75	75	75	74	74	74	74	72	72	71	70	69	68	68	68	68	67
ENERGY	82	80	82	80	81	81	81	80	81	80	79	81	81	77	78	80	76	74	74	78	73	76	73	75	72	76	72	73	72	67	71	69	70	65	67	69
WATER	79	80	79	78	78	77	78	78	79	78	78	76	78	77	76	75	75	75	74	74	75	71	71	71	74	75	70	72	70	70	69	68	69	69	65	67
DYNAMICS	82	82	82	84	83	82	81	82	80	82	82	81	77	81	81	76	78	79	77	72	76	77	79	77	76	72	75	69	71	71	67	68	66	69	72	64
ANNUAL	87	86	87	89	88	86	87	87	87	88	88	86	85	88	88	83	86	86	83	79	86	86	87	86	85	81	86	78	82	83	77	81	84	83	81	76
SEASONAL	88	88	88	89	88	85	86	87	86	86	88	85	85	85	88	83	86	86	82	81	85	87	87	87	84	83	86	80	84	84	78	83	74	84	81	78
ENSO	68	69	68	64	65	70	67	66	67	66	62	67	66	62	58	66	57	54	58	65	53	50	49	50	52	59	44	57	47	41	53	41	47	35	41	46
PRW	93	95	93	91	92	93	93	92	96	94	94	92	94	93	90	93	90	88	91	90	90	90	87	90	89	90	88	90	87	86	88	85	88	84	87	87
SLP	90	90	90	92	92	92	89	89	86	91	90	90	82	92	89	87	87	88	86	74	85	87	90	87	86	85	85	87	79	80	89	74	88	78	79	85
OLR	89	87	88	87	85	89	88	86	92	88	86	88	89	85	85	87	85	82	86	84	81	82	82	82	81	81	78	82	80	79	77	78	84	77	76	76
Z_{500}	87	86	87	90	91	89	85	89	82	85	87	88	80	87	85	83	83	85	79	79	77	83	87	83	81	79	80	81	79	76	81	74	82	74	73	74
RH_{500}	90	88	90	89	86	89	90	89	86	89	87	88	84	85	85	82	82	83	84	81	83	81	81	81	82	76	81	75	81	80	70	76	52	77	77	69
ASR	90	88	90	85	87	90	89	86	86	88	84	90	89	82	83	88	81	76	76	88	78	78	78	78	76	84	74	81	75	66	79	72	66	64	74	78
LH	84	85	84	85	85	78	82	84	81	83	82	79	80	82	81	80	79	78	76	75	80	79	80	80	79	80	75	79	77	73	80	75	76	72	71	76
U_S	79	84	79	87	81	79	79	79	78	81	83	79	79	82	81	72	79	79	79	78	74	80	84	79	74	79	80	79	83	69	77	79	80	70	76	76
LW^{CF}	83	77	82	81	79	81	82	80	87	82	79	80	81	79	79	77	75	74	75	75	74	75	67	74	73	71	69	69	68	70	63	67	77	68	56	63
R_T-F_S	78	79	78	77	77	76	78	79	76	78	75	75	76	72	75	77	75	74	74	75	74	75	74	76	73	74	74	74	74	71	74	71	71	70	68	70
P	80	80	79	78	79	79	80	80	82	77	78	78	78	77	78	74	75	75	73	74	69	65	67	66	68	75	64	69	64	71	64	62	61	69	65	61
E-P	79	80	79	78	79	76	76	78	81	77	79	75	77	76	79	74	71	75	71	73	71	63	65	63	70	72	59	69	64	65	64	63	62	63	62	62
SW^{CF}	80	77	80	76	81	81	78	76	75	71	78	81	78	74	77	79	71	69	67	80	66	72	67	71	65	74	66	63	65	55	61	62	53	52	67	60
F_S	75	75	75	73	74	71	73	75	71	74	70	70	71	68	70	73	70	68	66	69	66	71	71	71	65	70	70	68	68	63	72	63	68	61	61	68
RH_S	59	62	58	59	56	59	58	58	56	57	56	58	61	57	52	53	57	59	57	57	65	57	57	57	65	57	63	56	60	58	50	58	60	57	39	50
W^{500}	65	64	65	64	65	63	64	63	66	63	64	61	60	61	63	58	58	57	56	46	59	53	54	53	58	41	48	25	35	49	19	38	28	47	53	18

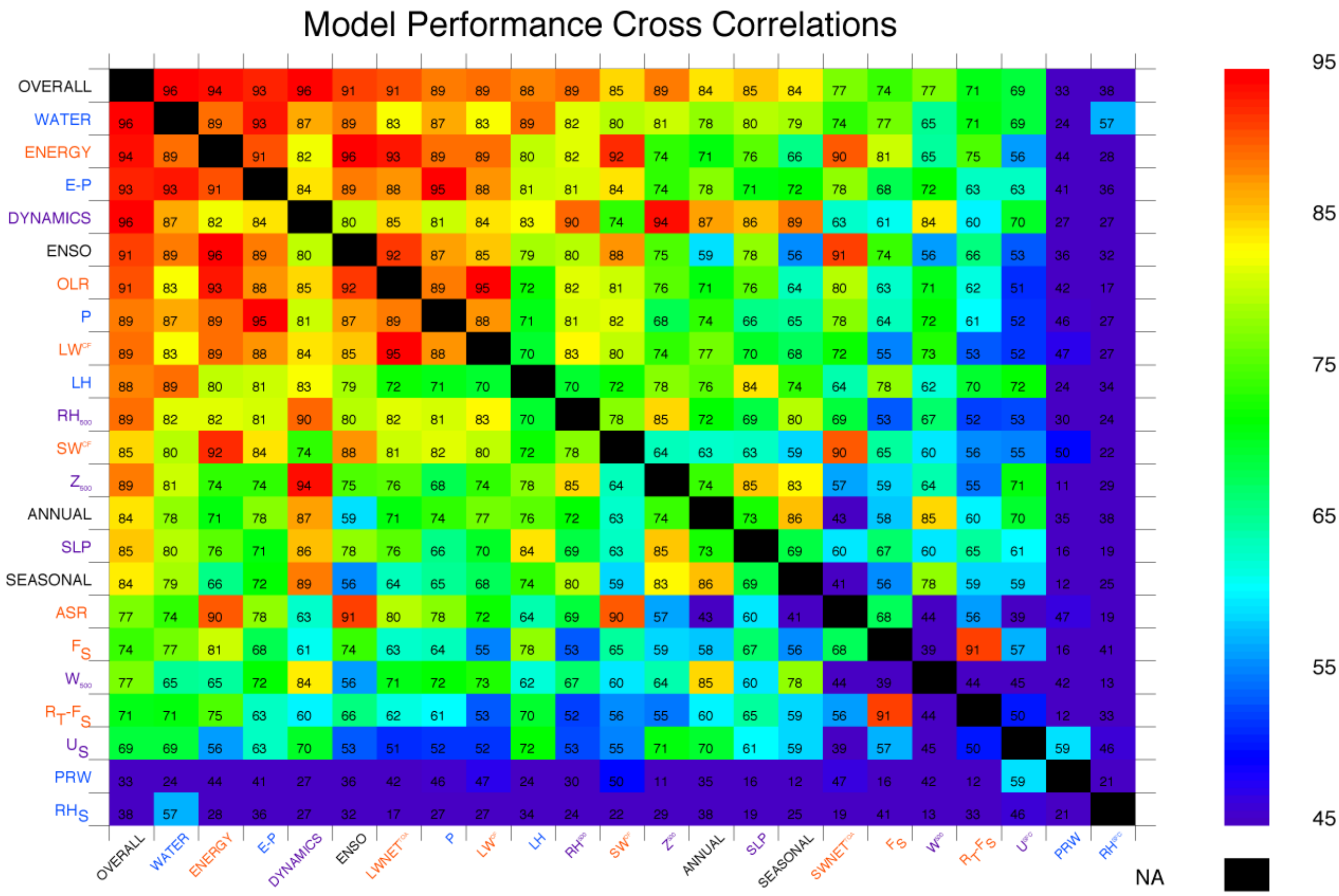
Figure 3: As in Fig. 2 except for historical simulations submitted to the CMIP5 archive.

https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020



	CESM2	MIROC6	SAM0-UNICON	CESM2-WACCM	GFDL-CM4	MRI-ESM2-0	CNRM-CM6-1	CNRM-ESM2-1	E3SM-1-0	EC-Earth3-Veg	CAMS-CSM1-0	GISS-E2-1-H	BCC-ESM1	CanESM5	BCC-CSM2-MR	UKESM1-0-LL	IPSL-CM6A-LR	GISS-E2-1-G	MIROC-ES2L
OVERALL	85	85	85	84	84	82	81	81	80	80	78	78	77	77	77	76	74	73	73
ENERGY	85	86	83	83	85	81	80	80	80	82	79	77	80	74	78	79	73	77	75
WATER	83	83	83	82	82	81	79	79	77	76	78	76	74	76	73	75	74	75	73
DYNAMICS	89	87	88	87	84	83	83	83	85	83	77	82	77	80	79	72	75	68	72
ANNUAL	91	89	91	91	91	90	88	88	88	87	84	89	85	88	86	81	86	81	80
SEASONAL	90	89	90	90	90	89	88	88	88	86	83	89	85	87	86	78	87	78	79
ENSO	74	77	72	71	70	66	65	66	65	68	67	57	61	55	57	68	49	61	60
PRW	96	95	95	95	96	93	94	94	95	94	92	90	92	90	92	93	90	90	91
SLP	95	93	93	94	89	90	91	89	93	91	84	92	89	89	91	88	89	84	86
RH_{500}	95	93	92	94	93	89	87	88	90	90	88	83	86	83	87	85	82	84	80
OLR	92	94	90	92	92	86	86	87	87	89	88	82	88	83	88	85	79	81	81
Z_{500}	92	90	91	91	85	87	87	89	89	86	75	88	84	87	84	82	85	76	84
ASR	89	92	89	86	89	85	87	87	85	88	85	80	86	77	81	89	74	85	85
U_S	90	87	91	89	88	82	85	84	87	83	78	87	68	84	75	82	83	77	77
LH	88	84	86	87	86	87	83	83	85	84	80	86	81	78	78	80	81	79	77
LW^{CF}	88	89	85	88	86	80	78	78	80	82	82	77	80	76	81	74	69	71	72
SW^{CF}	86	84	84	82	85	79	79	79	80	81	73	75	79	70	76	77	68	74	75
R_T-F_S	79	79	75	79	80	80	78	78	77	78	77	75	77	73	73	77	76	76	70
P	83	85	82	82	83	76	77	78	78	74	80	72	77	73	73	72	68	73	69
E-P	84	84	82	82	81	80	77	78	78	74	77	73	72	74	73	70	71	71	68
F_S	74	75	74	73	76	76	73	73	72	73	72	73	70	68	67	73	72	73	67
RH_S	64	65	69	63	64	70	66	63	49	55	62	58	50	67	48	63	62	63	58
W_{500}	71	69	74	70	64	66	63	64	64	65	61	60	59	57	59	24	36	20	32

Figure 4: As in Fig. 2 except for historical simulations submitted to the CMIP6 archive.



Model Performance Cross Correlations

	OVERALL	WATER	ENERGY	E-P	DYNAMICS	ENSO	$LWNET^{TOA}$	P	LW^{CF}	LH	RH^{500}	SW^{CF}	Z^{500}	ANNUAL	SLP	SEASONAL	$SWNET^{TOA}$	F_S	W^{500}	R_T-F_S	U^{SFC}	PRW	RH^{SFC}
OVERALL		96	94	93	96	91	91	89	89	88	89	85	89	84	85	84	77	74	77	71	69	33	38
WATER	96		89	93	87	89	83	87	83	89	82	80	81	78	80	79	74	77	65	71	69	24	57
ENERGY	94	89		91	82	96	93	89	89	80	82	92	74	71	76	66	90	81	65	75	56	44	28
E-P	93	93	91		84	89	88	95	88	81	81	84	74	78	71	72	78	68	72	63	63	41	36
DYNAMICS	96	87	82	84		80	85	81	84	83	90	74	94	87	86	89	63	61	84	60	70	27	27
ENSO	91	89	96	89	80		92	87	85	79	80	88	75	59	78	56	91	74	56	66	53	36	32
OLR	91	83	93	88	85	92		89	95	72	82	81	76	71	76	64	80	63	71	62	51	42	17
P	89	87	89	95	81	87	89		88	71	81	82	68	74	66	65	78	64	72	61	52	46	27
LW^{CF}	89	83	89	88	84	85	95	88		70	83	80	74	77	70	68	72	55	73	53	52	47	27
LH	88	89	80	81	83	79	72	71	70		70	72	78	76	84	74	64	78	62	70	72	24	34
RH_{500}	89	82	82	81	90	80	82	81	83	70		78	85	72	69	80	69	53	67	52	53	30	24
SW^{CF}	85	80	92	84	74	88	81	82	80	72	78		64	63	63	59	90	65	60	56	55	50	22
Z_{500}	89	81	74	74	94	75	76	68	74	78	85	64		74	85	83	57	59	64	55	71	11	29
ANNUAL	84	78	71	78	87	59	71	74	77	76	72	63	74		73	86	43	58	85	60	70	35	38
SLP	85	80	76	71	86	78	76	66	70	84	69	63	85	73		69	60	67	60	65	61	16	19
SEASONAL	84	79	66	72	89	56	64	65	68	74	80	59	83	86	69		41	56	78	59	59	12	25
ASR	77	74	90	78	63	91	80	78	72	64	69	90	57	43	60	41		68	44	56	39	47	19
F_S	74	77	81	68	61	74	63	64	55	78	53	65	59	58	67	56	68		39	91	57	16	41
W_{500}	77	65	65	72	84	56	71	72	73	62	67	60	64	85	60	78	44	39		44	45	42	13
R_T-F_S	71	71	75	63	60	66	62	61	53	70	52	56	55	60	65	59	56	91	44		50	12	33
U_S	69	69	56	63	70	53	51	52	52	72	53	55	71	70	61	59	39	57	45	50		59	46
PRW	33	24	44	41	27	36	42	46	47	24	30	50	11	35	16	12	47	16	42	12	59		21
RH_S	38	57	28	36	27	32	17	27	27	34	24	22	29	38	19	25	19	41	13	33	46	21	

430 **Figure 5: Cross correlations between variable and aggregate scores computed for the all CMIP archives sorted in order of decreasing correlations from left to right and top to bottom.**

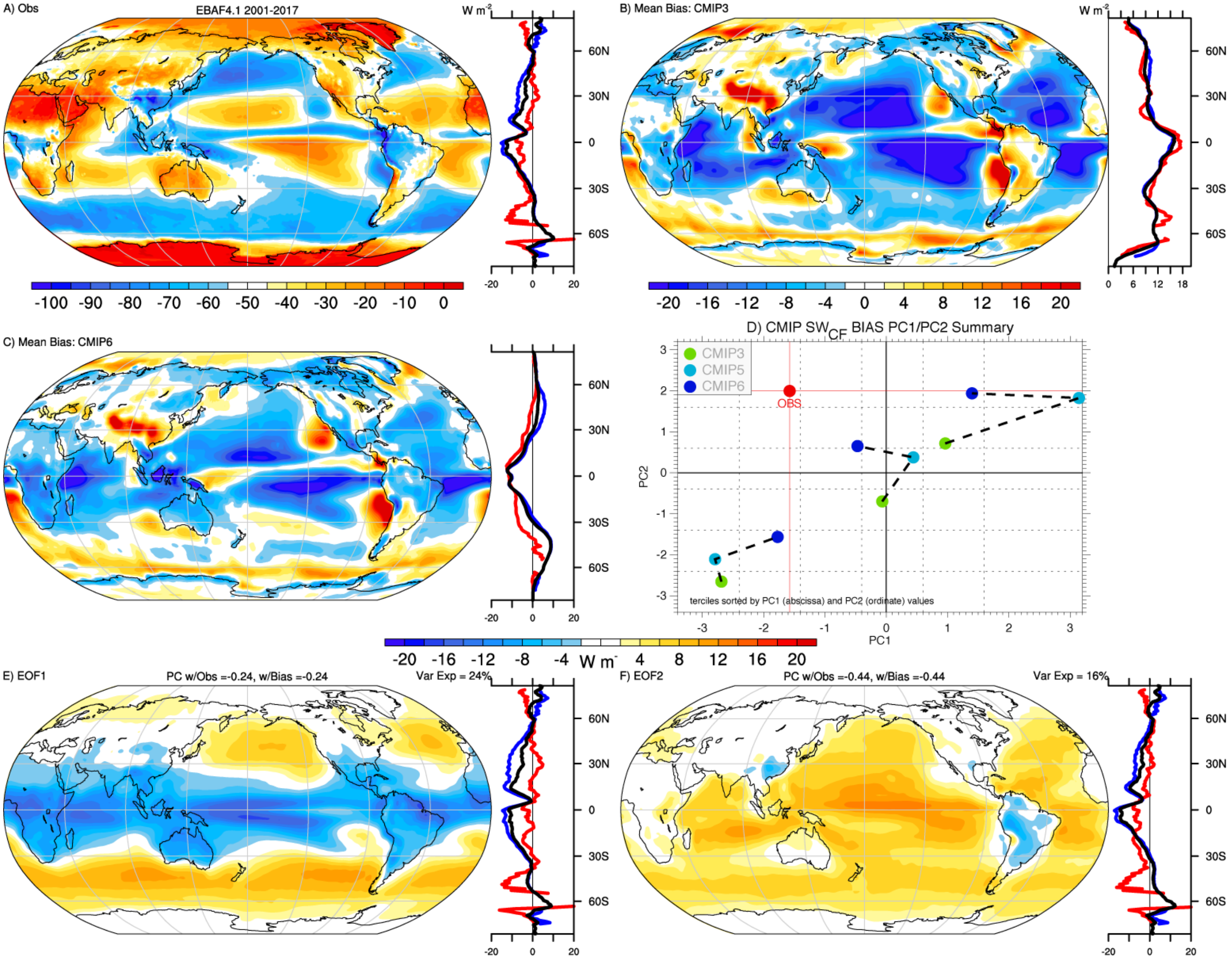


Figure 6: Analysis of the annual mean SW_{CF} bias in the combined historical CMIP3/5/6 archive including A) the observed estimate from CERES EBAFv4.1, the bias in (B) CMIP3 and (C) CMIP6, and (D) the first two PCs of biases and their tercile averages across the CMIP archives, and the associated first (E) and second (F) EOFs of biases. All units are W m^{-2}, except for the PCs, which are unitless.

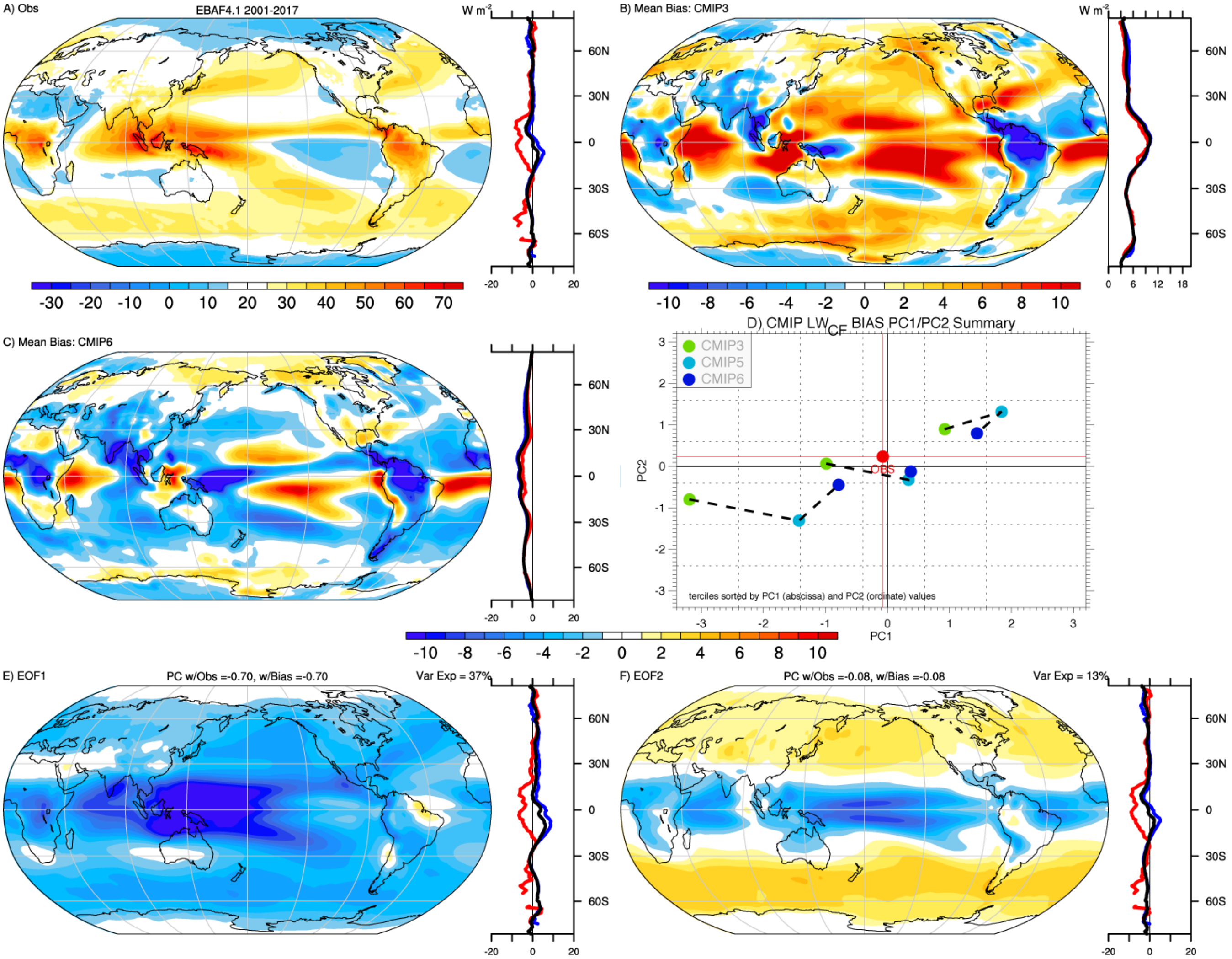


Figure 7: Analysis of the annual mean LW$_{CF}$ bias in the combined historical CMIP3/5/6 archive including A) the observed estimate from CERES EBAFv4.1, the bias in (B) CMIP3 and (C) CMIP6, and (D) the first two PCs of biases and their tercile averages across the CMIP archives, and the associated first (E) and second (F) EOFs of biases. All units are W m^{-2}, except for the PCs, which are unitless.

https://doi.org/10.5194/gmd-2020-1
Preprint. Discussion started: 28 February 2020

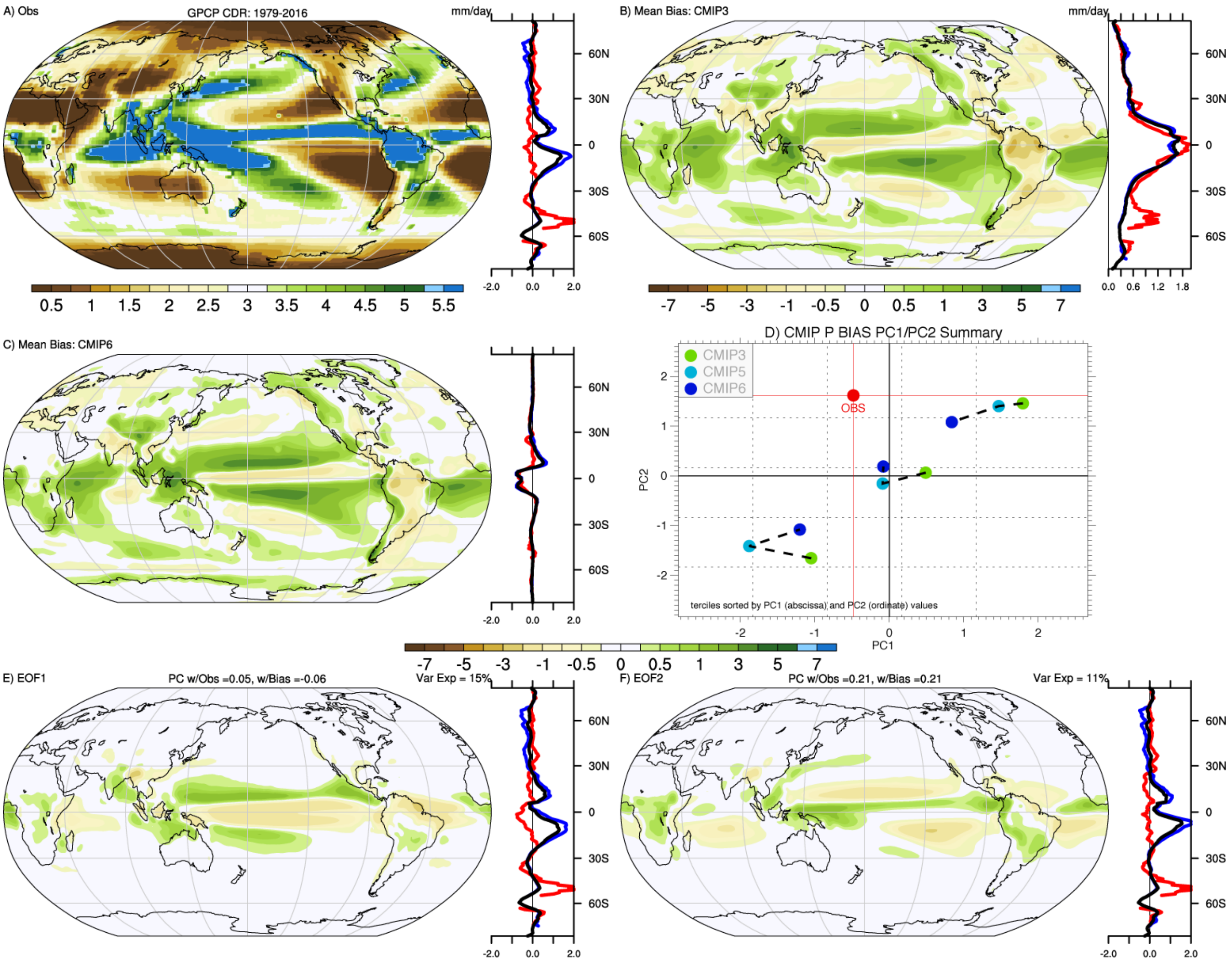


Figure 8: Analysis of the annual mean precipitation bias in the combined historical CMIP3/5/6 archive including A) the observed estimate from GPCP CDR, the bias in (B) CMIP3 and (C) CMIP6, and (D) the first two PCs of biases and their tercile averages across the CMIP archives, and the associated first (E) and second (F) EOFs of biases. All units are mm day^{-1}, except for the PCs, which are unitless.

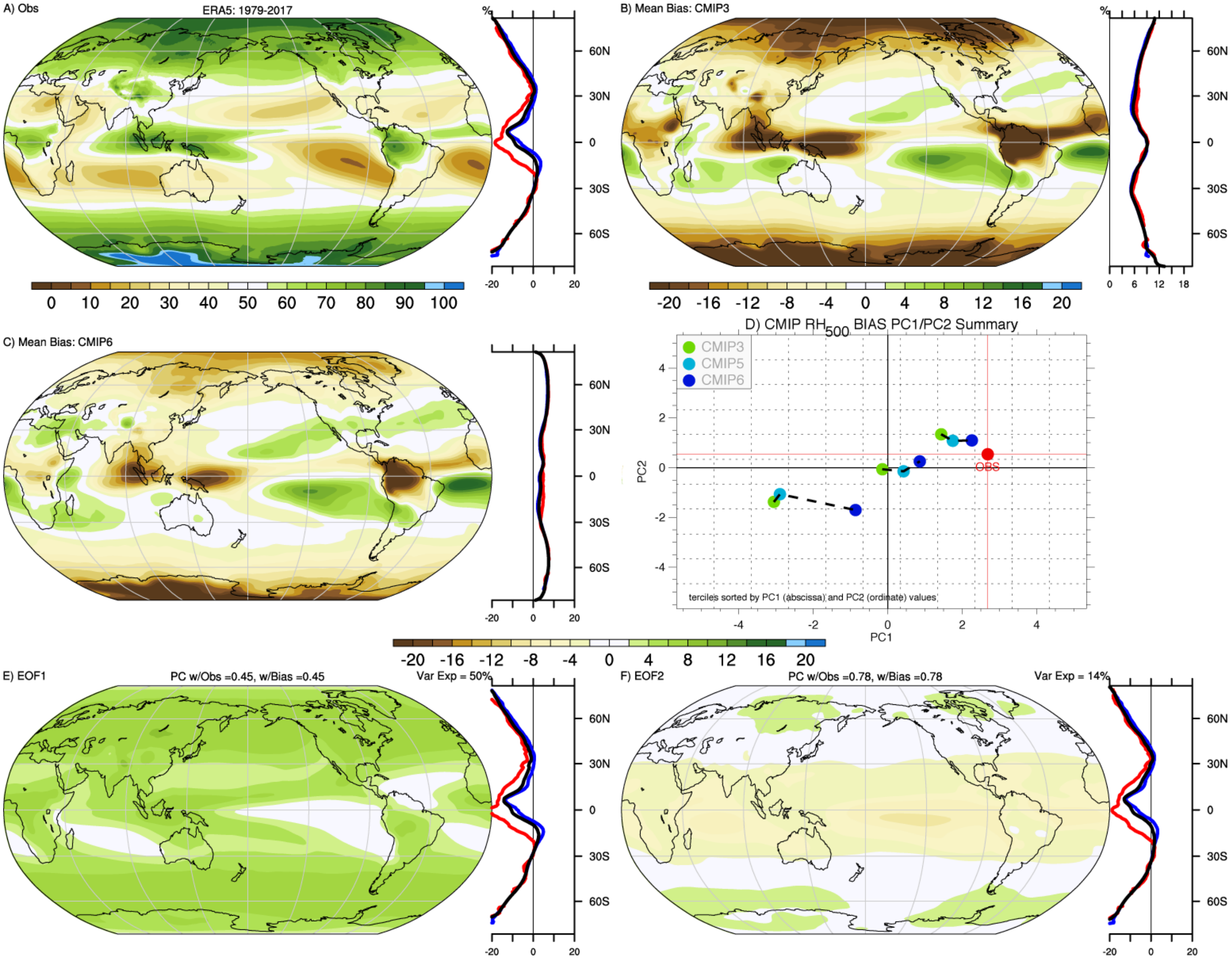
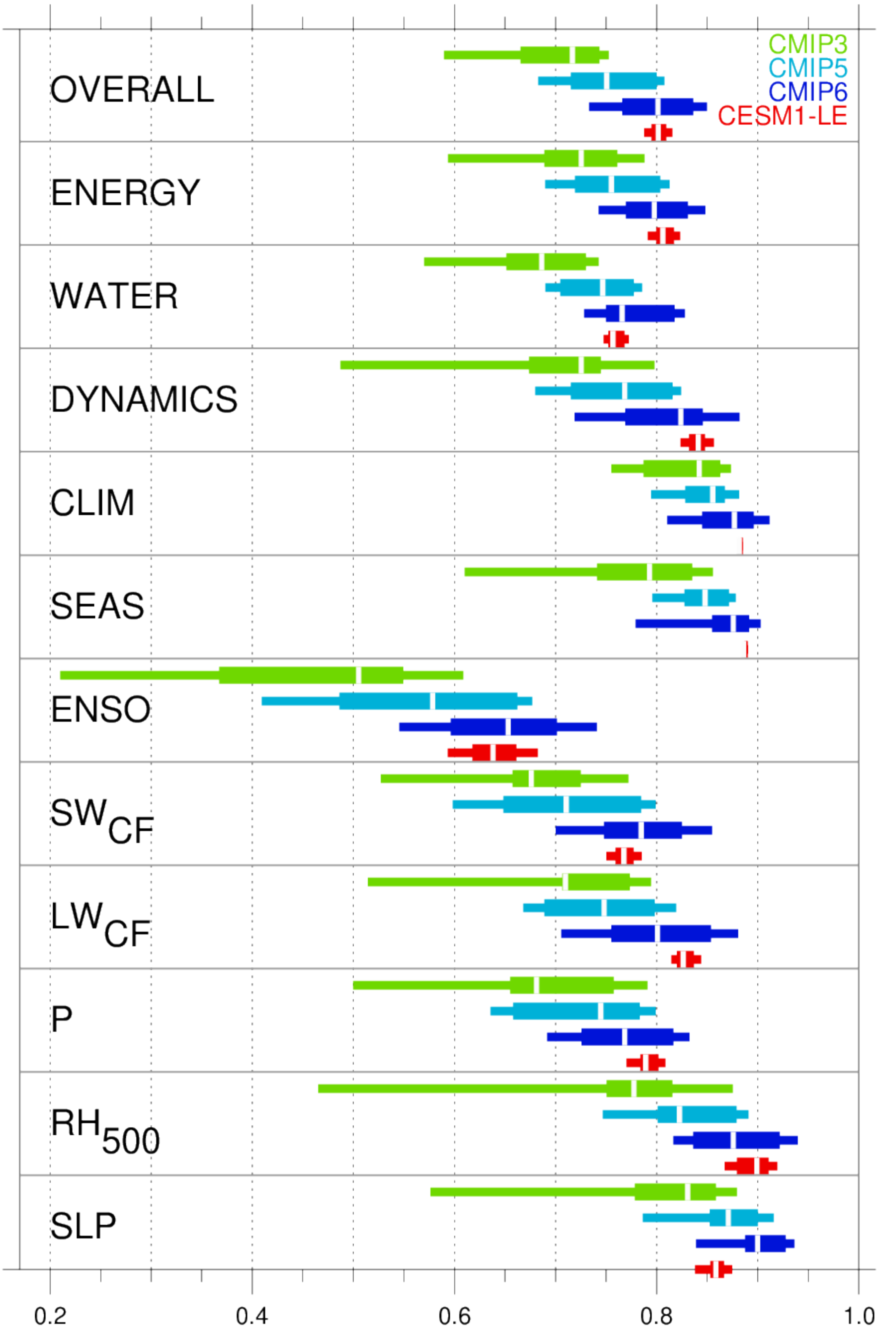


Figure 9: Analysis of the annual mean RH$_{500}$ bias in the combined historical CMIP3/5/6 archive including A) the observed estimate from ERA5, the bias in (B) CMIP3 and (C) CMIP6, and (D) the first two PCs of biases and their tercile averages across the CMIP archives, and the associated first (E) and second (F) EOFs of biases. All units are %, except for the PCs, which are unitless.



460 **Figure 10**: Evolution of the distribution of aggregate and selected variable scores across the CMIP archives and the CESM1-LE.