Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend

Yongxiao Liang, Nathan P. Gillett, & Adam H. Monahan

May 2020

Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL086757

Key Points:
- The trend in global mean temperature over 1970–2014 is well-correlated with projected future warming across the CMIP6 multimodel ensemble
- A weighting method based on the realism and degree of independence of simulated 1970–2014 trends predicts 2081–2100 warming well in cross-validation
- Observational constraints substantially reduce the upper bounds of projected 21st century warming ranges under the SSP scenarios

Supporting Information:
- Supporting Information S1

Correspondence to:
Y. Liang,
yongxiao@uvic.ca

Citation:
Liang, Y., Gillett, N. P., & Monahan, A. H. (2020). Climate model projections of 21st century global warming constrained using the observed warming trend. *Geophysical Research Letters*, *47*, e2019GL086757. https://doi.org/10.1029/2019GL086757

Received 25 DEC 2019
Accepted 4 MAY 2020
Accepted article online 10 MAY 2020

Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend

Yongxiao Liang[1], Nathan P. Gillett[2], and Adam H. Monahan[1]

[1]School of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia, Canada, [2]Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, Canada

Abstract The Coupled Model Intercomparison Project Phase 6 (CMIP6) archive includes larger ensembles, longer historical simulations, and models with a broader range of climate sensitivity than CMIP5. These features favor the application of observationally constrained climate projections. The 1970–2014 trend in global mean temperature is well-correlated with projected future warming across the CMIP6 multimodel ensemble. We first evaluate an approach that weights simulations based on the realism and degree of independence of their 1970–2014 trends, by treating each historical simulation in turn as pseudo-observations, and using the other models and weighting method to predict 21st century warming in the model concerned. The method performs well based on correlation and probabilistic measures. Applying the method using the observed 1970–2014 warming trend results in only small changes in the mean and lower bound of CMIP6 projected warming but substantially reduces the upper bound of projected early-, mid- and late-21st century warming under all SSP scenarios.

Plain Language Summary Different climate models predict different amounts of future warming over the 21st century. Most previous studies have weighted models equally to derive a range of projected future warming. Here, we test and apply an approach that gives more weight to models, which are better able to match the observed 1970–2014 warming trend. This approach substantially reduces the upper bound of projected warming over the 21st century.

1. Introduction

Uncertainties in climate model projections of future climate change result from the use of different emissions scenarios, model imperfections, and natural variability (Deser et al., 2012; Knutti et al., 2017; Knutti & Sedlacek, 2013). The Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) included a range of model projections of long-term warming without any performance-based weighting (Collins et al., 2014). Projections in the IPCC's Sixth Assessment Report (AR6) will be based largely on CMIP6 (Eyring et al., 2016). Compared to CMIP5, the number of different models, model variants, and ensemble sizes of individual models have all increased in CMIP6. Future scenario simulations in CMIP6 were coordinated by the ScenarioMIP project (O'Neill et al., 2016) and are driven by a new set of emissions and land use scenarios, known as Shared Socioeconomic Pathways (SSPs) (Riahi et al., 2017), produced using scenarios of future socioeconomic development to drive integrated assessment models. Some new CMIP6 models show higher transient climate response (TCR) and equilibrium climate sensitivity (ECS) compared with previous versions of these models in CMIP5 (Gettelman et al., 2019; Sellar et al., 2019; Swart et al., 2019; Voldoire et al., 2019; Zelinka et al., 2020), with some models warming more strongly than observations in recent decades (Swart et al., 2019). Multiple studies have argued for approaches other than using an unweighted ensemble of climate models to make projections, as not all models are equally skillful in reproducing observations (Brunner et al., 2019; Gillett, 2015; Knutti et al., 2017; Lorenz et al., 2018). In addition, CMIP6 includes multiple versions of similar models with differing resolution or differing model components. It may not be appropriate to use the arithmetic multimodel mean across all models given that the multimodel ensemble includes multiple closely related versions of some models, which are not independent (Knutti et al., 2013; Masson & Knutti, 2011). Here, we apply a weighting method defined by Knutti et al. (2017), following Sanderson et al. (2015a, 2015b, 2017). This method weights climate model simulations based on performance and independence

(Brunner et al., 2019; Knutti et al., 2017; Lorenz et al., 2018; Sanderson et al., 2015a, 2015b). Knutti et al. (2017) weighted multimodel projections of Arctic sea ice and temperature based on meaures of sea ice and temperature mean state, variability, and trends, showing that the weighting reduces model spread and projects a more rapid sea ice decline than the unweighted ensemble.

A recent study by Jimenez-de-la-Cuesta and Mauritsen (2019) showed that warming of indidvidual CMIP5 models over the post-1970 period is highly correlated with their TCR and used this relationship to constrain the TCR. Similarly, Nijsse et al. (2020) find a strong correlation between post-1970s warming and TCR in the CMIP6 models. The strong relationship between post-1970 warming and TCR arises because the temperature change since 1970 has been dominated by the response to greenhouse gases, with only a small aerosol-induced temperature change over this period (Jimenez-de-la-Cuesta & Mauritsen, 2019; Nijsse et al., 2020). The aerosol cooling exhibited an increase up until around 1970, with strong differences in forcing and response between models, meaning that temperature changes since the preindustrial period are not as strongly correlated with TCR across models (Forster et al., 2013; Jimenez-de-la-Cuesta & Mauritsen, 2019; Nijsse et al., 2020; Tokarska et al., 2020). Therefore, as the observational record continues to lengthen, observed warming trends for the post-1970s period may be a good metric to constrain projected warming. Recently, Tokarska et al. (2020) applied a regression-based approach to constrain future warming from CMIP6 simulations based on observed warming trends: Here, we apply a complementary weighting-based approach.

In this study, we apply the weighting approach of Knutti et al. (2017) to projections of 21st century warming, weighting simulations based on their historical temperature trends. In section 2 we describe the data sets and the methods used. Our results are shown in section 3. First, we explore the selection of a time period over which the simulated trend is well-correlated with projected future warming. Then we evaluate the weighting method and weighting metric using a cross-validated imperfect model test and probabilistic validation. This section also assesses the use of weights based on gridded surface air temperature and the implications of using all available ensemble members as opposed to using a single realization per model. We then present projections for changes in global near-surface air temperature constrained by observations. Section 4 contains a summary and conclusions.

2. Data and Methods

2.1. Global Climate Model Data From CMIP6

The CMIP6 archive includes output from global climate models from institutions around the world. Historical model simulations (1850–2014) and projections (2015–2100) of climate change under each of the Tier 1 SSP scenarios are used in this study (O'Neill et al., 2016; Riahi et al., 2017). All available model simulations, including multiple initial condition ensembles for individual models, are considered in our analysis. Thirty models with up to 50 ensemble members each are included in the analysis (see supporting information Table S1). Our study focuses on changes in monthly-mean global-mean near-surface air temperature (GSAT) in historical and future periods. In a sensitivity analysis, we also use gridded SAT climatologies over the period 1979–2014.

2.2. Observations

The HadCRUT4 data set consists of monthly historical instrumental temperature records, combining sea surface temperature data from the UK Met Office Hadley Centre with land surface air temperature records from the University of East Anglia Climate Research Unit (CRU) (Bridgman & Oliver, 2006). In our model analysis we generally consider globally complete GSAT rather than using blended near-surface air temperature over land and ice and SST over the ocean masked with observational coverage as in HadCRUT4 (GMST). We calculated both the multimodel mean of GSAT and simulated blended GSAT over land and ice and SST over ocean masked with HadCRUT4 coverage, in a subset of CMIP6 historical simulations. The multimodel mean ratio of 1970–2014 trends in GSAT versus GMST trends over the subset of models was 1.074. Therefore, we scaled the observed HadCRUT4 trend over this period by this ratio, in order to estimate the observed globally complete GSAT trend and then used this value when we derived weights based on the observations. In the sensitivity analysis, we use gridded SAT from ERA5 (C3S, 2017) to calculate the difference in simulated and reanalysis climatologies (Text S1).

2.3. Imperfect Model Test

In order to compare the performance of the weighting method compared with unweighted averages, we conduct a cross-validated imperfect model analysis of the CMIP6 simulations. In order that the calculation not be dominated by a small number of models with large ensemble size, we randomly pick out one ensemble member per model to act as "truth" (referred to as the "pseudo-observations"), and the weighting approach is applied using individual ensemble members from all other models to predict this "truth." We also consider probabilistic validation of the imperfect model test. To validate weighted projections, we noted in which quintile of the projection (0–20%, 20–40%, etc.) pseudo-observations lie for each projection, across all models. Ideally, 20% of the projections would lie in the first quintile, 20% in the second quintile, and so on. As a sensitivity analysis, all of these calculations are also repeated using all available ensemble members to derive projections but with equal weights applied to each model.

2.4. Weighting Method

The weighting method used in this study is described by Knutti et al. (2017) and is based on Sanderson et al. (2015a, 2015b). The weights w_i (defined for each ensemble member for each model) account for both model performance and interdependence:

$$w_i = \frac{e^{-\frac{D_i^2}{\sigma_D^2}}}{1 + \sum_{j \neq i}^{M} e^{-\frac{s_{ij}^2}{\sigma_s^2}}}. \quad (1)$$

In the numerator of equation 1, D_i is the distance of model i to observations. The parameter σ_D controls how strongly model performance is weighted. For large values of σ_D the weight given to each model is approximately equal. Small values of this parameter imply a more stringent constraint, putting most of the weight on a few models. In the denominator of equation 1, M is the number of simulations, S_{ij} is the distance between models i and j, and σ_s the parameter that controls how strongly models are penalized due to similarity to other models (Knutti et al., 2017; Lorenz et al., 2018). Both D_i and S_{ij} are evaluated here as absolute differences in 1970–2014 temperature trends (simulated and observed for D_i, pairs of simulated for S_{ij}) normalized by their median across models. The method we use to calculate σ_s (Text S1) is that proposed by Brunner et al. (2019), and the method we use to calculate the shape parameter σ_D is that described by Knutti et al. (2017) and Lorenz et al. (2018).

In the CMIP6 archive, for each SSP, only single ensemble members are available for some models, while a large number of ensemble members are available for others. Using all available simulations in determining the uncertainty range in the imperfect model test could result in the models with large ensemble numbers dominating, even when model independence is taken into account in the weighting procedure. This fact motivates investigating how the uncertainty range is affected by considering the entire set of simulations or only considering single ensemble members from each model. Our analysis focuses on results with weights based on one randomly seleted ensemble member per model, with the random selection process repeated 5,000 times (note that models with small ensembles will less effectively sample the range of internal variability). For each random set of realizations, we determine σ_D as described in Text S2. As a sensitivity analysis, we also conduct an analysis using all ensemble members, giving equal weights to each ensemble member from individual models and calculating weights based on the ensemble mean for each model and further weighting individual ensemble members by the inverse of the ensemble size (Text S2). In this sensitivity analysis, the ensemble mean is used in the distance measure D_i in order to reduce the influence of internal variability. In a second sensitivity test using quantities other than the historical trend to weight models, we further consider weighting based on the root-mean-square-difference (RMSD) between historical and simulated gridded SAT and a compound metric, which combines temperature trend and RMSD of gridded SAT with equal weight in the distance metric. The specific steps to calculate RMSD of gridded SAT are presented in Text S3. Finally, in order to assess the performance of the weighting method, we compare results with those obtained by giving equal weight to each model.

3. Results

3.1. Selection of Time Period

The primary quantity that we use to weight models is the GSAT trend over the historical period. The trend in global mean temperature over 1970–2014 is correlated well with projected future warming across the CMIP6 multimodel ensemble (Text S4; Figure 1a; the correlation coefficient is 0.80). This correlation has the highest such correlation out of a range of time periods we considered (Text S4). The fact that this period results in a higher correlation than for example 1960–2014 is probably a consquence of the fact that aerosol forcing has not changed much over the 1970–2014 period, so most of the GSAT trend over this period will be driven by GHGs, similar to future changes (Jimenez-de-la-Cuesta & Mauritsen, 2019; Nijsse et al., 2020).

We also evaluate the correlation coefficient between the GSAT trend and future warming, randomly sampling one ensemble member per model (and repeating this process 5,000 times; Text S4). The corner plot of Figure 1a shows the distribution of correlation values. The correlation is always relatively high, with a mean value of 0.68, and 5–95% ranges of 0.57–0.75 for future periods of 2081–2100. The relationship between historical GSAT trend and future warming among models is robust.

Figure 1b shows a comparison of simulated and observed trends over 1970–2014 across the different CMIP6 models. Many models show higher GSAT trends compared with observations, associated with the high climate sensitivity of these models (Gettelman et al., 2019; Sellar et al., 2019; Swart et al., 2019; Voldoire et al., 2019). In particular, the GSAT trends of CanESM5, UKESM1, and CESM-WACCM show trends that are significantly higher than observed GSAT trends at the 5% level, based on a t test.

3.2. Evaluation of the Weighting Method

3.2.1. Imperfect Model Test

In order to evaluate the performance of the weighting scheme based on the GSAT trend, we use an imperfect model test of weighted and unweighted results from the CMIP6 historical simulations and future projections of 2041–2060 and 2081–2100 using one member per model, selected at random (Text S5). The performances of the unweighted approach and weighting scheme are compared with the simulation being used as pseudo-observations using root-mean-square-error (RMSE) and correlation (r) between pseudo observations and statistical model predictions (Figure 2). This procedure is repeated for 5,000 random samples from the full set of model simulations. As shown in Figures 2a and 2b, the weighted results using the GSAT trend show better performance than the unweighted results both for the historical period and future projections, as measured by both correlation and RMSE. The mean RMSE difference across the 5,000 samples is 0.004 K/y for the historical trend and 0.04 and 0.12 K, respectively, for for projected changes in 2041–2060 and 2081–2100). Compared with unweighted projections, the weighting method results in robustly large and positive correlation coefficients between pseudo observations and mean predicted warming (Figure 2c). While the correlation coefficient for the unweighted averages is large, it is negative (always close to −1 for both historical and future periods). This is because when a model with stronger than average trends is treated as pseudo-observations, the ensemble of remaining models will tend to have weaker trends than average across the full ensemble and vice versa. It is evident from Figure 2 that, based on the centred correlation coefficient, the unweighted average has essentially no skill at predicting the pseudo-observations, as expected. Repeating the imperfect model test using ensemble means results in larger correlation values (Figure S2, Table S4). This result is strongly influenced by the small number of models with high climate sensitivity and large ensembles.

Repeating these calculations using the RMSD of gridded SAT (Text S6), we find that this quantity is not as useful as the GSAT trend for constraining future warming with historical records. Therefore, we focus on the GSAT trend as our primary metric to apply the observational constraint.

3.2.2. Probabilistic Validation

Probabilistic validation of our approach is important, since we are concerned with whether our uncertainty estimates are robust. For probabilistic validation (Text S7), we noted in which quintile of the projection pseudo-observations lie for each weighted or unweighted projection, across all models, and then constructed a histogram of the relative frequencies for each quintile. As before, random samples of individual ensemble members are taken from each model, and the process is repeated 5,000 times.

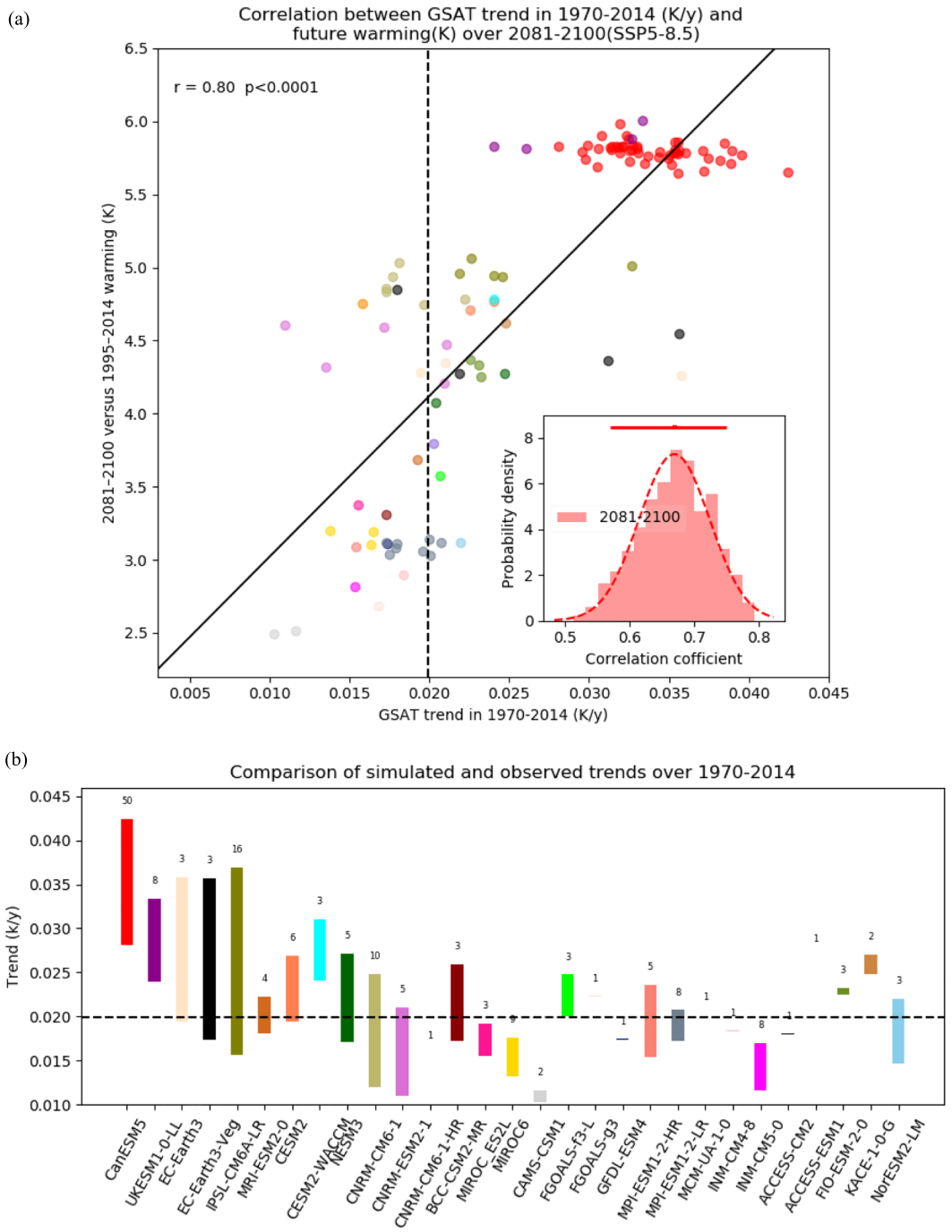

Figure 1. (a) Scatterplot of projected 2081–2100 warming relative to 1995–2014 under the SSP5-8.5 scenario against simulated 1970–2014 trends in GSAT. Colors correspond to those used in panel (b). Inset: probability density function (PDF) for correlation cofficient between GSAT trend and future warming based on 5,000 random samples of one ensemble member per model. The red histogram shows the PDF for correlation of historical GSAT trend and future warming in 2081–2100. The horizontal red line shows the corresponding 5–95% range, and the vertical tick shows the mean. (b) Comparison of simulated (colored bars) and observed (black dashed line) GSAT trends (units: K/y) over 1970–2014. The bars show uncertainty range for all model's ensemble members. The numbers marked at the bottom of each bar for panel b represent number of member in each model.

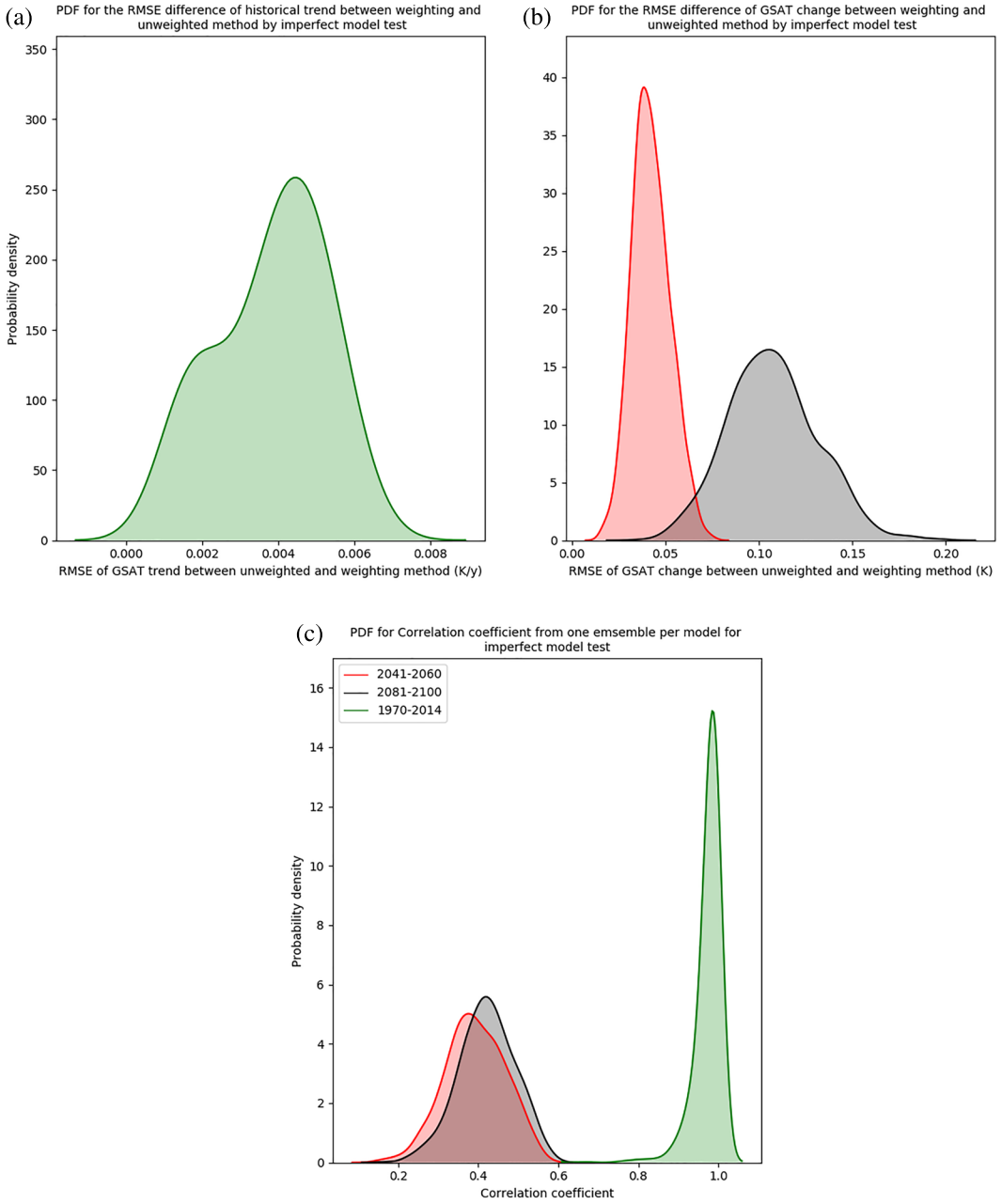

Figure 2. Reductions in RMSE due to the application of the weighting approach and correlations between mean weighted projections and pseudo-observations based on an imperfect model test with one ensemble member randomly stelected per model (repeated 5,000 times). Panel (a) and panel (b) show the distributions of RMSE decrease by weighting (relative to unweighted) for historical GSAT trends (green shading) and projected GSAT change under SSP5-8.5 (2041–2060 with red shading and 2081–2100 with black shading, respectively). Panel (c) shows the PDF of correlation coefficients for historical and future periods. We calculate correlation coefficients between the pseudo-observations and predicted means (both weighted and unweighted) for each random single-member per model sample. The red, black, and green shading show the correlation coefficients of weighted predicted means versus pseudo observation for 2041–2060, 2081–2100, and 1970–2014. The mean estimated correlation coefficient is 0.97 ($P < .01$ for all 5,000 samples) for the historical period, 0.40 (94% of 5,000 samples show $P < .1$) for 2041–2060, and 0.42 (98% of 5,000 samples show $P < .1$) for 2081–2100. The correlation coefficients of unweighted predicted means versus pseudo observation for all periods are always close to −1.

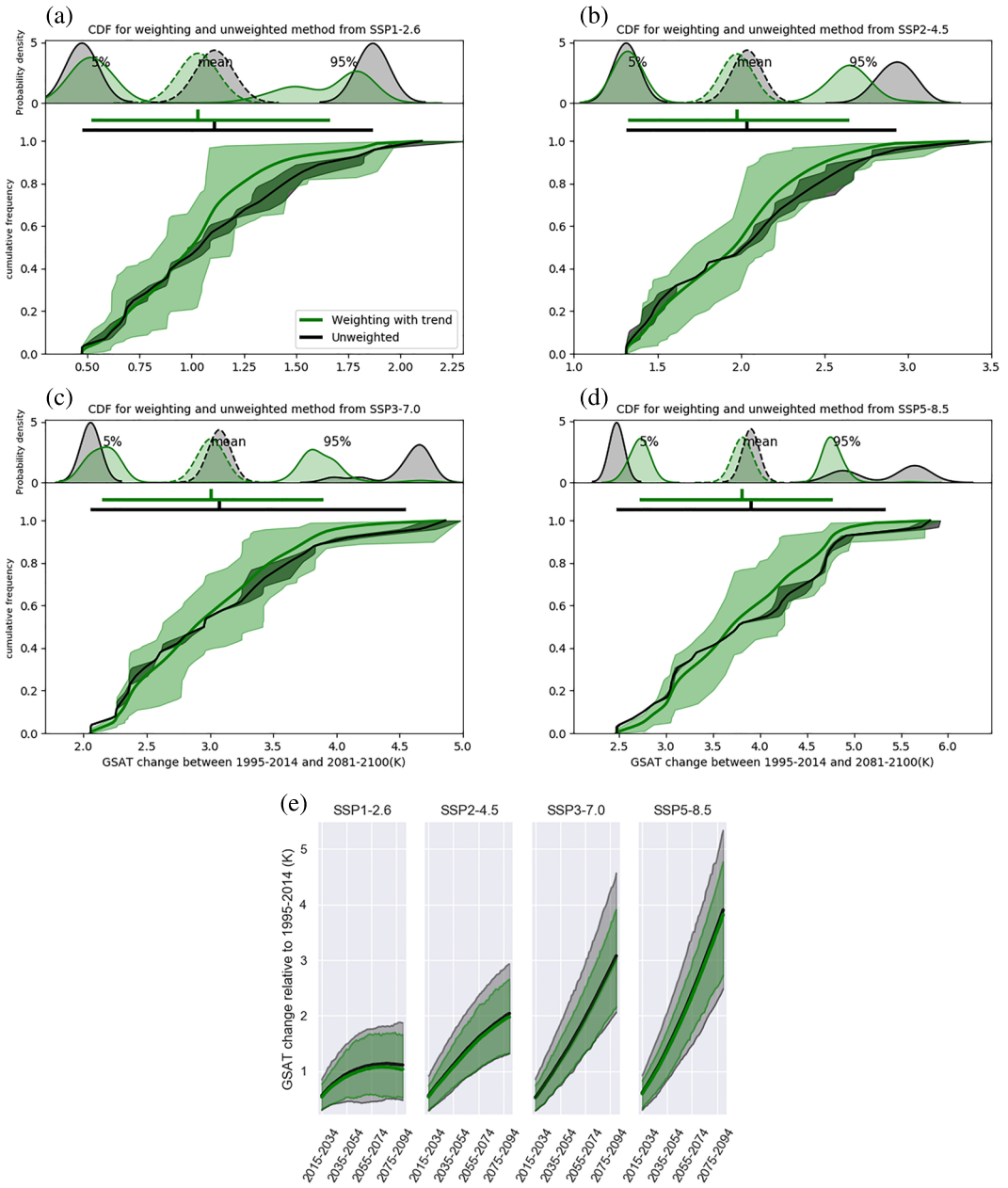

Figure 3. Distributions of projected GSAT warming between 1995–2014 and 2081–2100 in each of four scenarios (panels a–d), both constrained by observations (green) and unconstrained (black), based on 5,000 samples each with one randomly selected ensemble member per model. Unconstrained projections (black) are obtained giving equal weights to each model. The weights for the weighted method (green) are calculated based on the corresponding models' historical GSAT trends. The solid lines in green and black, respectively, represent the sample mean CDF for the weighted and unweighted method, respectively. Horizontal green and black lines show the best estimates of corresponding 5–95% ranges, and the vertical ticks show the corresponding means. The upper parts of panels (a–d) show the PDF of the 5th percentile, mean (dashed), and 95th percentile based on the distributions of projected GSAT warming between 1995–2014 and 2081–2100 in each of four scenarios. Panel (e) shows the best estimates of the 5–95% ranges of weighted (green) and unweighted (gray) results for other projection periods. The green (black) tick marks show the corresponding means of weighted (unweighted) results.

Figures S4a and S4b show the results of the probabilistic validation applied to mid-century and end-of-century projections respectively under SSP5-8.5. Even though the 5–95% ranges of weighted distributions are 25% narrower, the weighting approach gives approximately equal relative frequencies in each quintile, similar to the unweighted prediction. Note that the unweighted prediction, in which one of 29 models is

withheld and the CDF is constructed based on the remaining models with equal weight given to each model, is expected to perform well on this metric: If the validation was performed on the full ensemble without withholding models, the relative frequency for each quintile would be identically 0.2. We also applied the probabilistic validation using the full ensembles calculation (Text S7). Qualtitatively similar results were obtained (Figure S5).

3.3. Projections Constrained by Observations

Since the imperfect model analyses demonstrate that the weighting method has better performance than unweighted averages and probabilistic validation demonstrates that the weighting method performs well on uncertainty estimation, the weights obtained from the observed and simulated GSAT trends over the historical period by equation 1 can be applied to climate change projections for which we do not have observational estimates. Figures 3a–3d reveal that when using this weighting scheme, the distribution of weighted GSAT changes is narrower than that obtained without weighting for the projection period 2081–2100. This reduction of spread by weighting results occurs in other time periods as well (Figure 3e). Table S6 provides the 5–95% range and mean projection values for the four scenarios in each of three periods (2021–2040, 2041–2060, and 2081–2100), all computed as the average of the 5,000 single ensemble member samples. The weighted distribution of GSAT changes has a slightly lower mean than the unweighted model mean. For example, the weighted (unweighted) distribution has mean 3.82 K (3.90 K) in the high emission scenario SSP5-8.5 and 1.03 K (1.11 K) in the low emission scenario SSP1-2.6 in 2081–2100. The lower bounds of the projected ranges increase in the weighted ensemble, particularly for SSP5-8.5. However, the largest effect of the weighting is seen on the upper bound: The 95th percentiles of warming estimated from the CDF of the observationally constrained distribution (upper bound of green shaded band in Figure 3e or values in Table S6) are substantially lower than the corresponding unweighted values (upper bound of black shaded band in Figure 3e or values in Table S6) across all scenarios and periods. Finally, for SSP3-7.0 and SSP5-8.5, we find that the projection upper bounds show wide distributions across individual ensemble samples (Figures 3c and 3d); the widths of the distributions are reduced substantially by weighting.

We also constrain projections using all ensemble members by calculating weights by equation 1 based on the ensemble mean for each model (Text S2) and give equal weights to each ensemble member of a model. We find there is no substantial difference between the weighting results in this case and the means of the single ensemble samples (Table S6, Figure S6). When climate change projections are made using the compound metric involving GSAT trend and RMSD of gridded SAT, we find the results are close to the projection using GSAT trend alone (Table S6). As well, the model weights obtained from a compound metric involving both the GSAT trend and gridded SAT are close to weights obtained from the GSAT trend (Figure S7).

4. Summary and Conclusions

Consistent with the results of Forster et al. (2013), we find that projected warming in the CMIP6 simulations is not strongly correlated with warming over the full historical period, likely due to differences in aerosol forcing between models, and for this reason historical warming trends have not generally been used as a constraint on projected future warming. However, with the aerosol forcing and response having remained approximately constant since the 1970s (Forster et al., 2013; Nijsse et al., 2020), the lengthening observational record now affords us a period of more than four decades in which the observed climate response has been dominated by the effects of greenhouse gas increases and over which warming trends are closely correlated with Transient Climate Response (Daines et al., 2016; Jimenez-de-la-Cuesta & Mauritsen, 2019; Nijsse et al., 2020) and future warming in scenarios in which radiative forcing is dominated by further greenhouse gas increases. Hence in this study, we evaluate and apply an existing weighting method based both on model quality and independence (Knutti et al., 2017), to constrain projected warming in the CMIP6 simulations under the SSP scenarios using the GSAT trend.

Because of marked differences in the number of ensemble members provided for different models in the CMIP6 ensemble, we compared the results of weighting using ensemble means with the distribution of weighting results obtained by randomly sampling individual members from each ensemble. We find appreciable differences between values from the full-ensemble calculations and the mean values across the single ensemble member samples in imperfect model test results and hence focus our analysis on the latter

measure, which is less sensitive to differences in ensemble size across models. Nonetheless, differences in projected warming constrained by observations between the two approaches are small.

Since an assumption for using this weighting method to constrain projections is that models that have a realistic historical simulation also have realistic future projections (Brunner et al., 2019; Knutti et al., 2017; Lorenz et al., 2018), we evaluate the weighting method in an imperfect model test and compare with unweighted results. In the imperfect model test applied to mid-century warming (end-of-century warming) under SSP5-8.5 and considering means across 5,000 single-member per model samples, we find that the method gives 26% (25%) narrower best estimate confidence limits than the unweighted ensemble, with a correlation coefficient of 0.40 (0.42) between the mean weighted projection and truth and good performance in terms of probabilistic validation.

We also consider an RMSD-based metric and compound metric including GSAT trend and RMSD of gridded SAT. The metric based on RMSD in gridded SAT was not found to significantly improve the projections of mean warming compared to unweighted results in the imperfect model test. The compound metric was found to perform similarly to the trend-based metric. This result indicates that the RMSD of gridded SAT does not produce a robust constraint on future warming.

Applying the method to projected warming using the observed 1970–2014 GSAT trend as a constraint, we find lower mean projected warming under all scenarios and substantially lower 95th percentile warming in all cases. For example, we find best-estimate observationally constrained 5–95% warming ranges of 2.72–4.77 K and 0.52–1.66 K for 2081–2100 under the SSP5-8.5 and SSP1-2.6 scenarios, respectively, with upper bounds substantially lower than the corresponding unconstrained ranges of 2.48–5.34 K and 0.47–1.87 K for 2081–2100. For the 2021–2040 period, we find best-estimate observationally constrained 5–95% warming ranges of 0.48–1.06 K and 0.39–0.95 K under the SSP5-8.5 and SSP1-2.6 scenarios, respectively, also with upper bounds substantially lower than the corresponding unconstrained ranges of 0.43–1.25 K and 0.38–1.08 K. For the large-forcing scenarios SSP3-7.0 and SSP-8.5, the range of the 95th percentile warming across single-member per model samples is substantially reduced by weighting, relative to the unweighted range.

Acknowledgments
We thank World Climate Research Programme's Working Group on Coupled Modeling and the model output (http://esgf.llnl.gov) listed in Table S1 for their available simulations. We also thank John C. Fyfe and two anonymous reviewers for their helpful comments. We acknowledge Lukas Brunner and Ruth Lorenz for publishing their weighting code. This research was supported by the Discovery Grants Program of the Natural Sciences and Engineering Research Council of Canada.

References

Bridgman, H. A., & Oliver, J. E. (2006). Global climate system: Patterns, processes, and teleconnections. *Global Climate System: Patterns, Processes, and Teleconnections*, 1–331. doi:https://doi.org/10.1017/cbo9780511817984

Brunner, L., Lorenz, R., Zumwald, M., & Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, *14*(12). https://doi.org/10.1088/1748-9326/ab492f

C3S (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. date of access. https://cds.climate.copernicus.eu/cdsapp#!/home (accessed March 2019).

Collins, M., Knutti, R., Arblaster, J., Dufresne, J. L., Fichefet, T., Friedlingstein, P., et al. (2014). Long-term climate change: Projections, commitments and irreversibility. In T. F. Stocker, et al. (Eds.), *Climate Change 2013: the Physical Science Basis*, (pp. 1029–1136). Cambridge: Cambridge University Press.

Daines, J. T., Monahan, A. H., & Curry, C. L. (2016). Model-based projections and uncertainties of near-surface wind climate in Western Canada. *Journal of Applied Meteorology and Climatology*, *55*(10), 2229–2245. https://doi.org/10.1175/jamc-d-16-0091.1

Deser, C., Knutti, R., Solomon, S., & Phillips, A. S. (2012). Communication of the role of natural variability in future North American climate. *Nature Climate Change*, *2*(11), 775–779. https://doi.org/10.1038/nclimate1562

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

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

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*, 8329–8337. https://doi.org/10.1029/2019gl083978

Gillett, N. P. (2015). Weighting climate model projections using observational constraints. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, *373*(2054), 8. https://doi.org/10.1098/rsta.2014.0425

Jimenez-de-la-Cuesta, D., & Mauritsen, T. (2019). Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming. *Nature Geoscience*, *12*(11), 902. https://doi.org/10.1038/s41561-019-0463-y

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*, 1194–1199. https://doi.org/10.1002/grl.50256

Knutti, R., & Sedlacek, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, *3*(4), 369–373. https://doi.org/10.1038/nclimate1716

Knutti, R., Sedlacek, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*, 1909–1918. https://doi.org/10.1002/2016gl072012

Lorenz, R., Herger, N., Sedlacek, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *Journal of Geophysical Research: Atmospheres*, *123*, 4509–4526. https://doi.org/10.1029/2017jd027992

Masson, D., & Knutti, R. (2011). Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *Journal of Climate*, *24*(11), 2680–2692. https://doi.org/10.1175/2011jcli3513.1

Nijsse, F. J. M. M., Cox, P. M., & Williamson, M. S. (2020). An emergent constraint on Transient Climate Response from simulated historical warming in CMIP6 models. *Earth System Dynamics Discussions, 2020*, 1–14. https://doi.org/10.5194/esd-2019-86

O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., et al. (2017). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change-Human and Policy Dimensions*, *42*, 153–168. https://doi.org/10.1016/j.gloenvcha.2016.05.009

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015a). A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, *28*(13), 5171–5194. https://doi.org/10.1175/jcli-d-14-00362.1

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015b). Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate*, *28*(13), 5150–5170. https://doi.org/10.1175/jcli-d-14-00361.1

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, *10*(6), 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., et al. (2019). UKESM1: Description and evaluation of the UK Earth System Model. *Journal of Advances in Modeling Earth Systems*, *11*, 4513–4558. https://doi.org/10.1029/2019ms001739

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian Earth System Model version 5 (CanESM5.0.3). In (Vol. 2019, pp. 1--68). Geoscientific Model Development Discussions.

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). https://doi.org/10.1126/sciadv.aaz9549

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 DECK experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*, 2177–2213. https://doi.org/10.1029/2019ms001683

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*, 1–12. https://doi.org/10.1029/2019gl085782