Constraining Climate Model Projections of 21st-Century Global and Regional Warming

by

Yongxiao Liang

BSc, Nanjing University of Information Science & Technology, 2016

MSc, Nanjing University of Information Science & Technology, 2019

A Dissertation Submitted in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the

School of Earth and Ocean Sciences

© Yongxiao Liang, 2023

University of Victoria

All rights reserved. This dissertation may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.

Constraining Climate Model Projections of 21st-Century Global and Regional Warming

by

Yongxiao Liang

BSc, Nanjing University of Information Science & Technology, 2016

MSc, Nanjing University of Information Science & Technology, 2019

Supervisory Committee

Dr Nathan P. Gillett (Co-Supervisor)

School of Earth and Ocean Sciences, University of Victoria

Dr Adam H. Monahan (Co-Supervisor)

School of Earth and Ocean Sciences, University of Victoria

Dr Francis W. Zwiers (Outside Member)

Department of Mathematics and Statistics, University of Victoria

Dr Madeleine McPherson (Additional Member)

Department of Civil Engineering, University of Victoria

Abstract

Different climate models predict different amounts of future warming over the 21st century. Such uncertainty of future warming projections can be narrowed down by emergent constraints identified based on the relationships between projected warming across climate models and observable features of simulated past climate or climate change.

For global means of projected 21st-century warming, using the observed historical global mean near-surface air temperature (GSAT) trend as a constraint results in a relatively low warming relative to unconstrained projections. Using climatological cloud metrics, robust historical predictors with reduced influence of internal variability, to constrain future warming produces a relatively high warming. Such different ranges of constrained projections can be likely explained by the influence of internal variability in the constraint. By removing the unforced internal variability in historical GSAT trends, this study identifies a relatively higher 21st-century warming range than a constrained projection based on the raw GSAT trend, and brings GSAT trend constrained projections into much closer agreement with projections constrained using climatological cloud metrics. Regarding regional constraint of projected 21st-century warming, this study demonstrates the skill of global metrics relative to regional ones, and justifies the climatology cloud metrics alone can robustly constrain regional warming over extratropical Northern Hemisphere.

Contents

Supervisory Committee....ii
Abstract....iii
Contents....iv
List of Tables....vii
List of Figures....ix
List of Acronyms....xviii
Acknowledgements....xix
DEDICATION....xx
Chapter 1. Introduction....1
1.1 Constraining uncertainty of future global mean warming using the past warming trend....2
1.2. The sensitivity of constrained projected warming to different historical predictors....2
1.3. Constraining climate model responses with observations over regional scales....3
1.4. The reasons for lower future constrained global mean warming by applying the historical warming trend as a constraint....3
1.5. Structure of this dissertation....3
Chapter 2. Climate model projections of 21st century global warming constrained using the observed warming trend....5
2.1 Introduction and motivation....5
2.2 Data and methods....6
2.2.1 Global climate model data from CMIP6....6
2.2.2 Observations....7
2.2.3 Imperfect model test....7
2.2.4 Weighting Method....7
2.3 Results....8
2.3.1 Selection of time period....8
2.3.2 Evaluation of the weighting method....11
2.4 Summary and Conclusions....17
Chapter 3. Emergent Constraints on CMIP6 Climate Warming Projections: Contrasting Cloud- and Surface Temperature-Based Constraints....19
3.1 Introduction and motivation....19
3.2 Data and methods....22
3.2.1 Model simulations....22
3.2. 2 Emergent constraint metrics considered....23
3.2.3 Linear regression and step-wise metric selection....25
3.2.4 Sampling from initial condition ensembles and uncertain observational values....26
3.2.5 Imperfect model test....26
3.3 Results....27
3.3.1 Metric performance....27
3.3.2 Step-wise regression....28
3.3.3 Imperfect model evaluation of constrained warming....34

3.3.4 Observational contraints 36
3.4 Summary and discussion 40
Chapter 4. Observationally-constrained projections of 21st century regional warming in the extratropical Northern Hemisphere.....42
4.1 Introduction and motivation.....42
4.2 Data and Methods 44
4.3 Results.....50
4.4 Summary and discussion 58
Chapter 5. Constraining uncertainties in projected warming using the past global warming trend with the pattern effect removed 60
5.1 Introduction and motivation.....60
5.2 Data and methods.....61
5.2.1 Model data and observation 61
5.2.2 Constrained uncertainty estimates and imperfect model test for evaluation 62
5.2.3 Removing the unforced internal variability due to ETP SST trend from GSAT trend.....63
5.3 Results.....63
5.3.1 Observed and simulated ETP SST pattern 63
5.3.2 GSAT trend with the impact of the unforced ETP internal variability removed65
5.3.3 Performance of GSAT trend as a constraint.....66
5.3.4 Observationally constrained future GSAT changes 69
5.4 Discussion and Conclusion.....70
Chapter 6. Summary and Conclusions 73
6.1 Summary and Significance of Key Findings.....73
6.1.1 Chapter 2: Summary and significance of key findings.....73
6.1.2 Chapter 3: Summary and significance of key findings.....73
6.1.3 Chapter 4: Summary and significance of key findings.....74
6.1.4 Chapter 5: Summary and significance of key findings.....74
6.2. Synthesis of Results and Future Directions.....75
6.2.1 Synthesis of Results.....75
6.2.2 Future directions 76
Bibliography.....77
Appendix A.....83
Appendix B.....97
2.1 Materials and methods.....97
2.1.1 Linear regression method.....97
2.1.2 Weighting method.....98
2.1.3 Model dependence.....99
2.2 Supplementary information 100
Appendix C.....108
3.1 Materials and methods.....108
3.1.1 Metric selection strategy.....108
3.1.1.1 Stepwise selection process.....108

3.1.1.2 Lasso (least absolute shrinkage and selection operator) regression approach 108
3.1.2 Weighting approaches 109
3.2 Supplementary information 109
Appendix D.....116

List of Tables

Table 3. 1 CMIP6 Historical, SSP1-2.6, and SSP5-8.5 simulations used in this study. The number of ensemble members provided for each forcing senario is indicated in the second through the fourth columns.....22
Table 3. 2 Best estimates and 5-95% uncertainty ranges of projected warming using SSP5-8.5 for GSAT changes between 1995-2014 and 2081-2100. When calculating constrained uncertainty, we use a value of 20 as independent model amount in the CMIP6 ensemble (Appendix 2.1.1.3).....40

Table 4. 1 List of CMIP6 Historical and SSP5-8.5; and CMIP5 Historical and RCP8.5 simulations used in this paper. The numbers of ensemble members used for each experiment are listed in the second, the third, the fifth and the sixth columns. We used all simulations for which the necessary model output was available.....44
Table 4. 2 Short descriptions of each metric used as a constraint in this study. The global metrics (labelled as G) are MBLC, BCS and GSAT trend, while the rest of the metrics (labelled as R) are regional. The regional metrics are defined as the climatology, trend and variability of the area-averaged quantities, as described in Section 4.2b. The global (except GSAT trend) and regional metrics are calculated using data from 1980 to 2005 for both CMIP5 and CMIP6 ensembles in imperfect model test.....46

Table AA.S1 CMIP6 model runs for each Shared Socioeconomic Pathway (SSP) used in this study. For future projections, a complete set of simulations is not available for all models.....83
Table AA.S2 Weighting model parameters (σd and σs) for GSAT trend (m1), root-mean-square-difference (RMSD) of gridded SAT (m2) and the compound metric (m3) for different periods used in this study, derived based on ensemble means. 84
Table AA.S3 Correlation (r) between historical trends and projected warming (2081-2100 versus 1995-2014) under SSP5-8.5 for different trend periods : 1850-2014, 1960-2014, 1970-2014, 1980-2014 and 1990-2014. All p values for correlations listed in the table are less than 0.0001.....84
Table AA.S4 Correlation (r) and root-mean-square-error (RMSE) from the imperfect model test computed, based on different metrics, for SSP5-8.5 in 2041-2060 and 2081-2100. Values outside brackets are computed using ensemble means from each model, while values in brackets are means across 5000 single-member per model random samples.....84
Table AA.S5 Correlation (r) between the mean weighted projection and truth based on imperfect model test for two different metrics for the period of 2041-2060 from SSP5-8.5. We randomly select one member per model to do the imperfect model test. The third column shows the results when we remove one duplicated model for models from a single institution.....85
Table AA.S6 Projected mean warming and 5-95% confidence ranges based on the weighting method and unweighted simulations for four SSP scenarios (units: K). The

results for random selection on weighted and unweighted are best estimates from the 5000 samples.....85

Table AB.2. S 1 As Table 3.2, but derived using a value for the number of statistical degrees of freedom equal to the total number of models. 100

Table AC.2. S 1 CMIP6 Historical, SSP1-2.6 and SSP5-8.5 simulations used in this study. The number of ensemble members provided for each forcing senario is indicated in the second through the fourth columns. 110
Table AC.2. S 2 Projected mean warming and 5-95% confidence ranges based on the constrained and unconstrained projections for two SSP scenarios over 2081-2100 relative to 1995-2014 (units: K).....111

Table AD.S1 List of CMIP6 Historical, SSP5-8.5 and SSP1-2.6 simulations used in this paper. The numbers of ensemble members used for each experiment are listed in the second, the second, the third and the fourth columns. We used all simulations for which the necessary model output was available. 117

List of Figures

Figure 2. 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 5000 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 (coloured 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. 10
Figure 2. 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 5000 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 verus pseudo observation for 2041-2060, 2081-2100 and 1970-2014. The mean estimated correlation coefficient is 0.97 ($P < 0.01$ for all 5000 samples) for the historical period, 0.40 (94% of 5000 samples show $P < 0.1$) for 2041-2060, and 0.42 (98% of 5000 samples show $P < 0.1$) for 2081-2100. The correlation coefficients of unweighted predicted means verus pseudo observation for all periods are always close to -1. 13
Figure 2. 3 Distributions of projected GSAT warming between 1995-2014 and 2081-2100 in each of four scenarios (Panel a-d), both constrained by observations (green) and unconstrained (black), based on 5000 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 (grey) results for other projection periods. The green (black) tick marks show the corresponding means of weighted (unweighted) results. 16

Figure 3. 1 Correlation coefficients between potential observational constraints and projected warming. 5-95% uncertainty range and mean of correlation coefficients between potential constraints, evaluated from historical simulations, and simulated warming in response to the SSP5-8.5 scenario in 2081-2100 (relative to the reference period 1995-2014) based on 10,000 random samples from the initial condition ensembles (Section 2.4). The *p*-values of the mean correlation coefficients for BCA, GT, MBLC, BCS and LTMI are 0.004, 0.0007, 0.005, 0.13 and 0.32, respectively. The horizontal grey line represents the correlation value that is significant at the 0.05 level with the number of degrees of freedom estimated based on the number of independent models (Appendix 1.3). For display purposes, the signs of the MBLC metric and BCA correlations have been reversed.....28
Figure 3. 2 A flow chart of the step-wise regression procedure using cloud metrics. For each step, the corresponding statistics shown in Fig 3.3.....30
Figure 3. 3 5-95% uncertainty range and mean of F statistics at each step in the step-wise regression including only cloud metrics. The horizonal dotted lines represent critical F values at the 0.1 level. As discussed in Section 3.2.4, the 5-95% uncertainty ranges are generated by randomly sampling from the initial condition ensembles. For step 2 to step 4 in Fig 3.3, the vertical lines represent the F statistics obtained taking a value of 20 as number of statistical degrees of freedom, based on an estimate of the number of independent models in the CMIP6 ensemble (Appendix 2.2.1.3, Text S1). 31
Figure 3. 4 Schematic plot showing the physical basis of BCS and MBLC metrics. For the BCS metric, models which have a stronger convective control of cloud cover in subsidence regions in the current climate tend to have shallower clouds, and tend to have a larger reduction in cloud cover associated with strengthened convective drying as the climate warms. For the MBLC metric, models with a larger decrease of MBLC fraction in response to the SST warming at seasonal scale tend to have a larger decrease of MBLC fraction to SST warming at the centennial scale. MBLC and BCS metrics focus respectively on mid-latitude and tropical low-level clouds. Both these metrics are calculated over subsidence regions over the ocean. Detailed definitions of these selected metrics are in Section 3.2.2.....32
Figure 3. 5 Scatter plots showing relationships between selected constraints and projected warming. GT, MBLC (x-axis reversed), and BCS metrics are respectively shown in panels a, b and c. For illustration, one ensemble member per model is used. The correlation coefficients and *p*-values (relative to a null hypothesis of no correlation) are reported in the bottom right corner of each panel. The vertical lines show the observational values with means in solid and standard deviation in shadow. The dashed lines in each panel show the 66% confidence interval of the linear regression model [Appendix 2.1 eq (5)-eq (7)].....34
Figure 3. 6 Evaluation of constraining approaches with stepwise selected constraints using an imperfect model test. Panel (a) shows correlations between predicted means of constrained projections and pseudo-observations. Panel (b) shows reductions in

RMSE of constrained projections compared to unconstrained projections. The performance of the liner regression model is shown in green bars and the performance of the weighting method is shown in red bars, using stepwise selected metrics. Please note that the 5-95% uncertainty ranges (vertical bars) and means (dots) in panel (a)-(b) are a result of initial condition sampling. 35
Figure 3. 7 Histograms show the relative frequency with which the true 21st century warming in the individual SSP5-8.5 simulations lies within each of five quintiles of projected warming derived using the unconstrained and MBLC+BCS constrained approaches in an imperfect model test, aggregated across all models. Bars denote the median of the 10,000 single-member per model samples. The ±1 standard deviation ranges are denoted in error bars for each quintile. Note that the constrained distributions are slightly narrower than that the unconstrained distributions. (b) The frequency of the fraction of pseudo-observations lying in the 5-95% constrained uncertainty range across 10,000 samples. The blue bars and red dot show the frequency and the mean of 10,000 samples, respectively. 36
Figure 3. 8 Schematic plot to illustrate how observationally-constrained projections of warming are obtained using the GSAT trend metric with our Monte Carlo approach. The scatter plot on the left shows projected warming against historical warming in individual CMIP6 simulations, with one ensemble member chosen at random from each model. Two representative random samples of ensemble members and observations are illustrated in red and green. The associated regression relation (solid line and associated dashed lines show the linear regression model with corresponding 90% prediction interval), together with a realization of the observations (vertical dashed line), sampled from within its uncertainty range, is used to infer a PDF of projected warming, as shown in the right panel. The process is repeated 10000 times, and the corresponding PDFs are averaged to obtain the constrained projection (refer to Appendix 2.2.1.1). 37
Figure 3. 9 Constrained 20-year moving average GSAT anomalies derived using the linear regression approach with each of the cloud metrics and with the GSAT trend, compared to observations (based period: 1961-1990). The observational record we use is HadCRUT5 (spatially infilled version). The x-axis shows the centre of the 20-year averaging period. The green (GSAT trend) and grey (cloud metrics: MBLC and BCS) shadows show 5-95% constrained uncertainty ranges with solid lines showing the best estimates. We account for internal variability and observational uncertainty by constructing the regression models using one randomly selected ensemble member per model and using observed quantities sampled from their uncertainty ranges (assuming Gaussian distributions with means and standard deviations quoted in section 3.2.2) and then repeating this 10,000 times. 38
Figure 3. 10 PDFs of constrained and unconstrained GSAT changes between 2081-2100 and 1995-2014 under SSP5-8.5. The bottom panel shows the predicted distribution of GSAT changes constrained using the GSAT trend (blue), constrained using cloud metrics (green) and the unconstrained distribution (black). The shadows around these PDF curves displays the contribution of internal variability and observational uncertainty, estimated by sampling one ensemble member per model

and sampling the observed quantities within their uncertainty ranges (assuming Gaussian distributions with means and standard deviations quoted in section 3.2.2) 10,000 times. The solid curves correspond to the mean of these 10,000 samples. The upper horizonal bars display the respective 5-95% projected ranges and means (numerical values are given in Table 3.2) corresponding to the mean solid curves (the theoretical basis for this calculation is shown in eq (3.8) of Appendix 2.1.1.1). These results are obtained assuming a value of 20 for the number of statistical degrees of freedom of the CMIP6 ensemble (Appendix 2.1.1.3).....39

Figure 4. 1 The regions used in this study (Iturbide et al. 2020). 44
Figure 4. 2 Correlation coefficient between historical predictors and future regional warming by applying bootstrap resampling over models and initial condition ensembles (1000 times). 5-95% uncertainty ranges are shown. The sign of the correlation coefficient of MBLC is reversed.....51
Figure 4. 3 Imperfect model test of the accuracy of the constrained temperature changes for 2081-2100 relative to 1995-2014 using SSP5-8.5 for CMIP6 and RCP 8.5 for CMIP5. The left panels show correlation coefficients calculated between pseudo-observations and projected means and the right panels show the RMSE reduction calculated as the RMSE of unconstrained means relative to pseudo-observations minus the RMSE of constrained means relative to pseudo-observations. a) and b) compare the regression and Sanderson weighting approaches using cloud metrics and CMIP6 data. b) and c) compare the performance of different sets of metrics using the regression approach and CMIP6 data, and e) and f) show the same comparison using CMIP5 data. Bars show the 5^{th}-95^{th} percentile range across 5000 samples, sampling across model ensemble members. The black dashed line in the left panel plots show the critical value of the correlation coefficient that is significant at the 0.05 level (one-sided) for CMIP5 and CMIP6 respectively. The black dashed line in the right panel plots shows the threshold value of 0 representing no improvement based on the metrics chosen. All color bars shown in the plot are based on the linear regression method described in Section 4.2e except the dark blue bars which are based on Sanderson weighting approach using MBLC and BCS constraints. 53
Figure 4. 4 Widths of constrained and unconstrained 5%-95% ensemble uncertainty ranges. The bars show 5^{th}-95^{th} percentile ranges of ensemble widths based on 5000 random selections of model ensemble members. 54
Figure 4. 5 Evaluation of the reliability of the uncertainty estimates on the constrained projections. The upper panel shows the percentage of pseudo-observations lying within constrained uncertainty ranges (or coverage ratio) for the in-sample test based on CMIP6 simulations and the lower panel shows corresponding results for the split sample test based on CMIP5 simulations. The bars show 5-95th percentile ranges based on 5000 samples of initial condition ensembles. The bars in black represent constrained projections using Lasso selected metrics and the bars in green represent constrained projections using cloud metrics. 55
Figure 4. 6 20-year moving average of regional mean near-surface air temperature anomalies (based period:1995-2014) in CMIP6 future projections (SSP5-8.5). The red

solid lines show the 5-95% uncertainty range of raw CMIP6 projections and green solid lines show the constrained uncertainty range. The dashed lines show the ensemble mean estimates of constrained and unconstrained projections. The times shown on the *x*-axis are the central years of each 20-year moving average starting from 2015-2034 and ends at 2081-2100.....56
Figure 4. 7 As in Fig 4.6, but for SSP1-2.6.....57
Figure 4. 8 The percentage decrease in the 5-95% uncertainty width due to application of the observational constraints. The constrained and unconstrained widths are derived from Fig. 4.6 and 7 for the 2081-2100 of the 21st century relative to 1995-2014. The upper panel is for SSP 5-8.5 and the lower panel is for SSP 1-2.6.....58

Figure 5. 1 Observed and simulated patterns of SST trends in 1993-2012. Panel a-c show patterns of SST trend in observation, the ensemble member 'r7ilp1f2' of MIROC-ES2L and the ensemble 'r10ilp1f1' of CESM2, respectively. Panel d shows the PDF of ETP trend using CMIP6 ensembles. The PDF is drawn by sampling one random ensemble per model 5000 times. The solid black curve is the mean of the 5000 samples while the grey shading shows the range of these 5000 samples. The colored dots in panel (d) correspond to Panel (a), (b) and (c) respectively. The horizonal bars show the ensemble range of trend for individual model CanESM5 and MIROC6. 65
Figure 5. 2 Removal of influence of unforced internal variability of ETP SST on GSAT trend. Panel (a): the relation between the raw GSAT trend and eastern tropical Pacific temperature trend with forced response removed. The black regression line is estimated from all model realizations (black dots). The black ellipse contains 90% of the probability of the joint distribution (assuming Gaussian distribution). Panel (b): the PDFs and observations (vertical line) of raw GSAT trend (blue) and of GSAT trend with ETP SST trend removed (red). 66
Figure 5. 3 Intermodel correlation between GSAT trend (based on 1970-2022) and projected warming using SSP 5-8.5 (based on 2081-2100 relative to 1995-2014). Panel (a) shows the correlation between historical GSAT trend and projected GSAT changes aross all models. Correlations are based on 5000 random selections of one ensemble per model. Panel (b) shows the intermodel correlation between GSAT trend and projected warming of each grid box (an average across 5000 random selections). The black dashed line in the left panel plot shows the critical value of correlation coefficient that is significant at the 0.05 level. The white shading in panel (b) represents correlation coefficients not significant at the 0.05 level. 67
Figure 5. 4 Imperfect model test of constrained projections. The raw simulations of warming change are based on 2081-2100 relative to 1995-2014 using SSP5-8.5 for CMIP6. Panel (a) plots distributions of correlation coefficient calculated by pseudo-observations and projected mean and Panel (b) plots distributions of RMSE reduction calculated by the RMSE of the unconstrained mean to pseudo-observations less the RMSE of constrained mean to pseudo-observations. The distributions in Panel (a) and (b) are based on 5000 random samples from ensembles (the bar show 5-95th percentile of 5000 samples). The black dashed line in panel (a) shows the critical value of

correlation coefficient that is significant at the 0.05 level. The black dashed line in panel (b) shows the threshold 0 representing no improvement from the constraint. Panel (c) shows the width of constrained uncertainty. The bars show 5-95th percentile range of 5000 random samples from the ensembles. Panel (d) shows the evaluation of the reliability of constrained uncertainty. The bars show 5-95th percentile of 5000 random draws from the ensembles.....68
Figure 5. 5 PDFs of constrained and unconstrained projected GSAT changes in 2081-2100 (SSP 5-8.5) relative to 1995-2014. The blue curve shows constrained projections using the raw GSAT trend, while the black curve uses the unconstrained projections. The red curve shows constrained projections using the GSAT trend with the ETP unforced pattern effect removed. The green curve shows projections constrained using climatological cloud metrics from Chapter 3. The distributions are generated by sampling over the internal variability as described in Section 5.2.2.....70

Figure AA.S1 (a) Inter-model RMSD in 1979-2014 mean gridded SAT (units: K). Each row and column represents a single climate model. Warm colors represent larger distance, while closer distances shown by cool colors. (b) Model and observation distance (units: K). Each bar represents a single climate model, the length of bar depend on the members' range. Models with large RMSD means far distance to observation, vice versa. The numbers marked at the bottom of each bar for panel b represent number of member in each model. 91
Figure AA.S2 Imperfect model test using ensemble means compared with the unweighted ensemble (green and black circles respectively) for historical (Panel a) and future periods of 2041-2060 (Panel b) and 2081-2100 (Panel c). In each plot, the x-axis represents pseudo observations and the y-axis represents the mean value predicted by the corresponding method. The orange line denotes the 1:1 line for which the predicted value is equal to pseudo-observations. The green and black lines are respectively linear least-squares fits for the weighting method and unweighted simulations. The values of correlation and RMSE for pseudo observations versus predicted means in this plot can be found in Table S4. 91
Figure AA.S3 Similar to Fig S2 but for the compound metric (red dots) and metric RMSD of gridded SAT (blue dots). The values of correlation and RMSE for pseudo observations versus predicted means in this plot can be found in Table S4. 92
Figure AA.S4 Probabilistic validation using one randomly-selected ensemble member per model. Histograms show the relative frequency with which the true 21st century warming in the individual SSP5-8.5 simulation lies within each of five quintiles of projected warming derived using the weighting method and unweighted approaches in an imperfect model test, aggregated across all models. Bars denote the median of the 5000 single-member per model samples. Figure S4a and Figure S4b show results for 2041-2060 and 2081-2100 respectively, relative to the 1995-2014 base period. The green and black bars correspond to the weighting method and unweighted model output. The error bars show +/-1 standard deviation ranges for each quintile. Note that the 5th-95th percentile range of the weighted distribution is about 25% smaller than that of the unweighted distribution.....93

Figure AA.S5 Probabilistic validation using ensemble means for the weighting. Histograms show the relative frequency with which true 21st century warming in the individual SSP5-8.5 simulation lies within each of five quintiles of projected warming derived using the weighting method and unweighted approaches in an imperfect model test, aggregated across all models. Relative frequency is weighted such that each model has equal weight irrespective of ensemble size. Figure S5a and Figure S5b show results for 2041-2060 and 2081-2100 warming respectively, relative to the 1995-2014 base period. Note that the 5th-95th percentile range of the weighted distribution is about 25% smaller than that of the unweighted distribution. 93
Figure AA.S6 Distributions of projected GSAT warming between 1995-2014 and 2081-2100 in each of four scenarios (Panel a-d), both constrained by observations (green) and unconstrained (black). Unconstrained projections are derived based on ensemble means of models, with equal weights given to each model. The weights for the weighting method are calculated based on the corresponding ensemble mean of models' historical GSAT trends, and then give equal weights to ensemble members. Horizontal green and black lines show the corresponding 5-95% ranges, and the vertical ticks show the corresponding means. Panel e shows 5-95% ranges of weighting (green shadow) and unweighted results (grey shadow) for other projection periods. The green (black) solid line show the corresponding means of weighting (unweighted) results.....94
Figure AA.S7 The models weights obtained by metric GSAT trend (circles) and compound metric (squares) from SSP5-8.5. Model weights are calculated by ensemble mean of each model. The orange line represents equal weights.....95

Figure AB.S1 Model-model distance matrix normalized by its median for gridded annual mean surface air temperature over the period of 1970-2014 (see details in Text S2 and Appendix 3.1.3).....102
Figure AB.S2 Model family tree for all 26 CMIP6 models used in this study. Models branching further to the right are more independent. Fig S2 is based on the model-model distance matrix shown in Fig S1 calculated from gridded near-surface air temperature. The dashed vertical line represents the independence shape parameter and is used to determine models that are independent or not (Brunner et al. 2020b). The gray shading represents an estimate of internal variability calculated from the median of distance between pairs of initial-condition realizations taken from the same model.....103
Figure AB.S3 Similar to Fig 3.3 but for F statistics calculated with a range of numbers of degrees of freedom. For each class of F statistic in the same color bar, the bars correspond to numbers of degrees of freedom ranging from the full number of models (topmost bar) to half the number of models (bottommost bar). The horizontal dashed lines represent the F statistics with effective degree of freedom determined from the models' genealogy (Appendix 3.1.3, Text S2), as also displayed in Figure 3.3. The vertical red shaded areas represent critical F values at the 0.1 level for the corresponding range of degrees of freedom.....104

Figure AB.S4 As in Fig S3 but with the stepwise selection based on low cloud and GSAT trend metrics. 105
Figure AB.S5 An assessment of how well the effects of internal variability are accounted for in our study, as described in Text S4. We first artificially set all ensemble sizes to one. We then select M models, with, M=1,2,3…12 in turn, from the set of models with ensemble size greater than one, and repeat the sampling calculation (described in Section 3.2.4) to get the distribution width (maximum minus minimum) of observationally constrained 5th (in black) and 95th (in red) percentiles. For each number of models, we show the mean across the sampling range. Solid lines correspond to MBLC and GSAT trend metrics, dashed lines to MBLC and BCS metrics. 106
Figure AB.S6 Similar to Fig 3.9, except using a number of statistical degrees of freedom equal to the number of models in the CMIP6 ensemble. 107
Figure AB.S7 Similar to Fig3. 9, but for SSP 1-2.6. 107

Figure AC.S1 Constrained warming in synthetic data experiment (described in Text S1). The PDFs are constrained projections for global mean surface air temperature change (2081-2100 relative to 1995-2014 under SSP 5-8.5) by applying MBLC and BCS over NWN (N.W.North-America). The dashed red PDF is derived using the approach used in the main manuscript, but sampling from synthetic 50-member ensembles from each model (see Step 2 of Text S1). The blue PDFs are derived from the same data, but first restricting ensemble sizes for each model to the number of simulations actually available (see Step 3 of Text S1). 112
Figure AC.S2 Same as Fig 4.2, but for CMIP5 ensembles. 113
Figure AC.S3 The predictors selected for each region based on the metric selection process. Panel (a) indicates predictors selected by the Lasso selection strategy and Panel (b) indicates predictors selected by the stepwise selection strategy. Predictors are described in Table 4.2. Panel (c) indicates predictors selected by Lasso selection strategy by potential predictors from Table S3 (the corresponding description of predictors to be chosen is in Text S2). Markers in red indicate global metrics while the markers in black, blue and green indicate trend-based, climatology-based and standard-deviation-based regional metrics respectively. 113
Figure AC.S4 As in Fig 4.3, but for CMIP6 SSP 1-2.6 using cloud metrics.....114
Figure AC.S5 As in Fig 4.4, but for CMIP6 SSP 1-2.6 using cloud metrics.....114
Figure AC.S6 As in Fig 4.5, but for CMIP6 SSP 1-2.6 using cloud metrics.....115

Figure AD.S1 Regress out the model difference on forced response regionally for ETP trend (based on 1970-2022). '*r*' represents the correlation coefficient between ECS and ETP trend across multi models with the corresponding '*p*' value, while '*N*' represent the number of climate models in use. 118
Figure AD.S2 Similar as Fig S1, but for the period of 1993-2012. 118
Figure AD.S3 Similar to Fig 5.2 but for the period of 1993-2012.....119
Figure AD.S4 Similar as Fig 5.3 but for the period of 1993-2012. 119
Figure AD.S5 Similar to Fig 5.5 but based on SSP 1-2.6. 120

Figure AD.S6 The correlation coefficient between historical GSAT trend (based on different initial years and the different final years) and future projected warming between 1995-2014 and 2081-2100 under SSP5-8.5. The *y-axis* represents the start years while *x-axis* represents the end years. Panel (a) shows the correlation coefficient between raw GSAT and projected warming while panel (b) shows the correlation coefficient between GSAT trend with ETP variability removed. Panel (c) shows the difference between panel (b) and (a). The shading areas in panel (c) represent two correlation coefficients that are significantly different from each other at level 0.05. 120

List of Acronyms

AR5	Fifth Assessment Report
AR6	Sixth Assessment Report
BCA	Brient Cloud Albedo (defined in Section 3.2.2)
BCS	Brient Cloud Shallowness (defined in Section 3.2.2)
CDF	Cumulative Distribution Function
CMIP6	Coupled Model Intercomparison Project Phase 6
CMIP5	Coupled Model Intercomparison Project Phase 5
ECS	Equilibrium Climate Sensitivity
ESS	Explained Sum of Squares
ETP	Eastern Tropical Pacific
GHG	Greenhouse Gas
GSAT	Global-mean Near-surface Air Temperature
GT	GSAT Trend
HFSS	Sensible Heat Flux at Surface
HFLS	Latent Heat Flux at Surface
IPCC	Intergovernmental Panel on Climate Change
LASSO	Least Absolute Shrinkage and Selection Operator
LTMI	Lower Tropospheric Mixing Index
MBLC	Marine Boundary Layer Cloud
OLR	Ordinary Least-squares Regression
PDF	Probability Density Function
PSL	Sea Level Pressure
PR	Precipitation
RMSD	Rroot-mean-square-difference
RMSE	Root-mean-square-error
RCP	Representative Concentration Pathway
RSDS	Surface Downwelling Shortwave Radiation
RLUS	Surface Upwelling Longwave Radiation
SSP	Shared Socio-economic Pathways
SST	Sea Surface Temperature
TCR	Transient Climate Response

Acknowledgements

We acknowledge the World Climate Research Programme's Working Group on the modeling of Coupled Model Intercomparison Project, and we thank the climate modelling groups for modelling and making their model output available.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the China Scholarship Council (CSC) and graduate fellowships from the School of Earth and Ocean Sciences at the University of Victoria, Canada.

I am extremely grateful to my supervisors: Dr Nathan Gillett and Dr Adam Monahan, for all of their guidance throughout this process.

I would also like to thank the committee members: Dr Nathan Gillett, Dr Adam Monahan, Dr Francis Zwiers, Dr Madeleine McPherson and Dr John Fyfe, provided very helpful comments and guidance for this dissertation.

I am also very grateful to Dr Xuebin Zhang for sharing his expertise and thoughtful advice.

Finally, I thank my family for their unwavering support throughout my academic career abroad, especially during the past 3 years of pandemic. I especially thank my grandpa, Zuolin, Liang, for his forever love.

DEDICATION

To my loving family, who always encourage me to strive for my dreams.

Chapter 1. Introduction

Climate projections are generally based on results from multiple climate or Earth System Model simulations [e.g. the Coupled Model Intercomparison Project Phase 6 (CMIP6), (Eyring et al. 2016)]. There are three main sources of uncertainty in model projections of climate change (Hawkins and Sutton 2009): that due to uncertainty in future anthropogenic forcings scenario uncertainty, due to internal climate variability, and due to uncertainty in the model difference of physical climate response. Projected future baseline emissions depend on the economic development processes, and societal choices (Riahi et al. 2017). Internal variability can be estimated by uncertainties from multiple initial condition ensembles of a single specific model. Multi-model ensembles can serve as a way of probing inter-model differences. For a specified emission scenario, the projection uncertainty based on multi-model ensembles is an obstacle to making confident and accurate future climate projections as models may not have equally plausible consistency with evidence from observations, theory, or process understanding (Caldwell et al. 2018; Jimenez-de-la-Cuesta; Mauritsen 2019; Nijsse et al. 2020; Tokarska et al. 2020). The multi-model mean may be biased relative to actual climate change when a number of models have systematic errors, or many closely related models are included. In addition, the spread in projected warming (e.g. in CMIP6) is too large to inform climate policies (Lee et al. 2021). Hence, it is important to narrow the raw spread of model ensembles. The so-called emergent constraints approach is designed to reduce the model spread by examining the collective behavior that emerges unexpectedly in climate model ensembles such as those assembled for the CMIP5/CMIP6. Specifically, emergent constraints rely on there being a physically explainable empirical relationship between an observable model metric over the historical period and future model projections (Hall and Qu 2005; Allen et al.2002). This connection is applied to constrain projections using a range of statistical approaches, including linear regression, weights, detection and attribution methods and Bayesian methods (Allen et al. 2000; Bindoff et al. 2014; Gillett et al. 2021; Stott and Kettleborough 2002; Stott et al. 2006; Brunner et al. 2020a; Brunner et al. 2020c; Knutti et al. 2017; Nijsse et al. 2020; Tokarska et al. 2020; Renoult et al. 2020; Ribes et al. 2021b; Ribes et al. 2022; Rougier et al. 2013).

This dissertation explores ways to constrain the uncertainty of future global mean warming (Chapter 2), assesses its sensitivity to different historical predictors (Chapter 3), investigates observationally constrained warming over regional scales (Chapter 4), and develops an interpretation of the relatively low constrained future global mean warming derived by constraining with the historical warming trend (Chapter 5). The specific research questions addressed are explicitly stated at the beginning of each chapter.

The following four sections in this chapter (Section 1.1- Section 1.4) serve as background information about the observationally-constrained future projections, and provide a brief review of the current literature relevant to this topic.

1.1 Constraining uncertainty of future global mean warming using the past warming trend

The global temperature change is approximately proportional to the radiative forcing and models with stronger climate feedbacks exhibit more warming in both the past and the future. Many studies rely on this concept to suggest future warming based on past warming (Jimenez-de-la-Cuesta and Mauritsen 2019; Nijsse et al. 2020; Tokarska et al. 2020). The relation between past and future warming is often obscured by time-varying climate feedbacks and uncertain aerosol forcing, especially in the historical period (Nijsse et al. 2020; Tokarska et al. 2020). However, this inter-model correlation (between past and future warming) has become significant over recent decades due to the greenhouse gas induced warming dominating the observed warming with a small aerosol-induced temperature change. Therefore, historical warming can theoretically serve as constraint to narrow the uncertainty of future warming, which is also dominated by greenhouse gases. For example, Nijsse et al (2020) find a robust correlation between post-1970s warming and transient climate response (TCR) in the CMIP6 and CMIP5 models. As the observational record continues to lengthen, observed warming trends for the post-1970s period could be a good metric to constrain projected warming. Recently, Tokarska et al. (2020) applied a regression-based approach to constrain a lower future warming relative to unconstrained projections from CMIP6 simulations using observed warming trends as a constraint. In Chapter 2, I apply a complementary weighting-based approach to constrain 21st century warming using the CMIP6 archive. My study evaluates the performance of a weighting method using a historical warming metric and compares with unweighted results. The results of Chapter 2 are published as Liang et al. (2020).

1.2. The sensitivity of constrained projected warming to different historical predictors

Model differences in simulated changes of shortwave reflection by low-level clouds in the tropics and midlatitudes dominate the uncertainties in equilibrium climate sensitivity (ECS) in past multi-model intercomparisons (Brient and Schneider 2016; Vial et al. 2013). Therefore, cloud related properties are promising for use as an emergent constraint. Caldwell et al. (2018) use a feedback decomposition analysis to evaluate the performance of several emergent constraints on ECS and show that only four of nineteen cloud constraints on ECS proposed in the literature can be considered physically credible.

As discussed in Section 1.1, recent analyses using the historical surface air temperature warming trend as a constraint favor low climate sensitivity models (Nijsse et al. 2020; Tokarska et al. 2020). However, higher climate sensitivity models are generally found to be in better agreement with observational constraints using cloud related metrics (Bretherton and Caldwell 2020; Brient and Schneider 2016; Caldwell et al. 2018; Zhai et al. 2015). In Chapter 3, I examine differences in projected warming based on these two constraints, and conduct a thorough comparison between cloud- and surface

temperature-based constraints in the observational constraint framework. The results of Chapter 3 are published as Liang et al. (2022)

1.3. Constraining climate model responses with observations over regional scales

Many studies show that large intermodel differences exist in modeling the magnitude of regional warming under climate change (Davy and Outten 2020; Lehner et al. 2020), but only a few attempts have been made to apply observational constraints at the regional scale where uncertainties are large (Brunner et al. 2019; Brunner et al. 2020a; Hegerl et al. 2021). However, Brunner et al. (2020c) and Hegerl et al. (2021) point out that diverse lines of evidence (e.g. different metrics as well as different approaches give different projections) lead to diverging constraints; and suggest that additional work is needed to understand how the underlying differences between methods lead to such disagreements.

In Chapter 4, I evaluate and constrain the projected surface air temperature averaged over sub-continental regions in the extratropical Northern Hemisphere, based on a set of potential constraints including global average climate metrics related to climate sensitivity as well as a series of regional climate metrics previously used in the literature. The results of Chapter 4 coresponding to Chapter 4 are submitted as Liang et al. (2023a).

1.4. The reasons for lower future constrained global mean warming by applying the historical warming trend as a constraint

There are some discrepancies in constrained projections with different metrics, e.g. higher climate sensitivity models are generally found to be in better agreement with observational constraints using cloud related metrics, while the historical surface air temperature warming trend finds the high sensitivity models to be less consistent and suggests that they might overestimate the future warming trend (Nijsse et al. 2020; Tokarska et al. 2020). However, constrained projections can be biased due to a strong influence of internal variability on the observed historical warming trend. In Chapter 5, we assess the influence of unforced internal variability on the past warming trend, and remove the influence of unforced internal variability in this predictor when using it as an emergent constraint. The results of Chapter 5 coresponding to Chapter 5 will be submitted as Liang et al. (2023b).

1.5. Structure of this dissertation

This dissertation explores the observationally constrained projections over global and regional scales, as specified in the Chapter 2-5.

Each of these four research areas is explained in more depth in the subsequent Chapters 2, 3, 4 and 5, respectively. Each chapter contains motivation, specific research questions, methods and preliminary results that are relevant for each project. Chapter 2-5 in this

dissertation investigated key aspects of constraining model uncertainty of projected warming, focusing on the performance of historical metrics as well as the role of internal variability in constrained projections. Notably, the understanding of the impact and importance of accounting for internal climate variability on the constrained projections evolved over the course of my research. Chapter2 reflects my understanding of internal variability at the time when Liang et al. (2020) was published, which subsequently evolved towards a view that gives internal climate variability much greater importance as in Chapter 5. The general conclusions are reported in Chapter 6.

Chapter 2. Climate model projections of 21st century global warming constrained using the observed warming trend

This chapter has been published as:

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(12), e2019GL086757.

2.1 Introduction and motivation

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 and Sedlacek 2013; Knutti et al. 2017). 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. 2020a), 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. 2019a; 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 2011b). 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. 2019a; 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 Jiminez-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 and 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 and 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 2.2 we describe the data sets and the methods used. Our results are shown in Section 2.3. 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 2.4 contains a summary and conclusions.

2.2 Data and methods

2.2.1 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 Table AA.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.

2.2.2 Observations

The HadCRUT4 dataset 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 and Oliver 2006; C3S 2017). 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 multi-model 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 multi-model 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 AA.S1).

2.2.3 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.

2.2.4 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:

$$w_i = \frac{e^{-\frac{Di^2}{\sigma_D^2}}}{1 + \sum_{j \neq i}^{M} e^{-\frac{s_{ij}{}^2}{\sigma_S{}^2}}} \tag{2.1}$$

In the numerator of Eqn (2.1), 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 Eqn (2.1), 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 AA.S1) is that proposed by Brunner et al. (2019a), 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 5000 times (note that models with small ensembles less effectively sample the range of internal variability). For each random set of realizations, we determine σ_D as described in Text S2 of Appendix 1. 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 AA.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.

2.3 Results

2.3.1 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 multi-model ensemble (Text S4 in Appendix 1; Figure 2.1a; the correlation coefficient is 0.80). This correlation has the highest such correlation out of a range of time periods we considered (Text S4 in

Appendix 1). 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 and Mauritsen 2019; Nijsse et al. 2020).

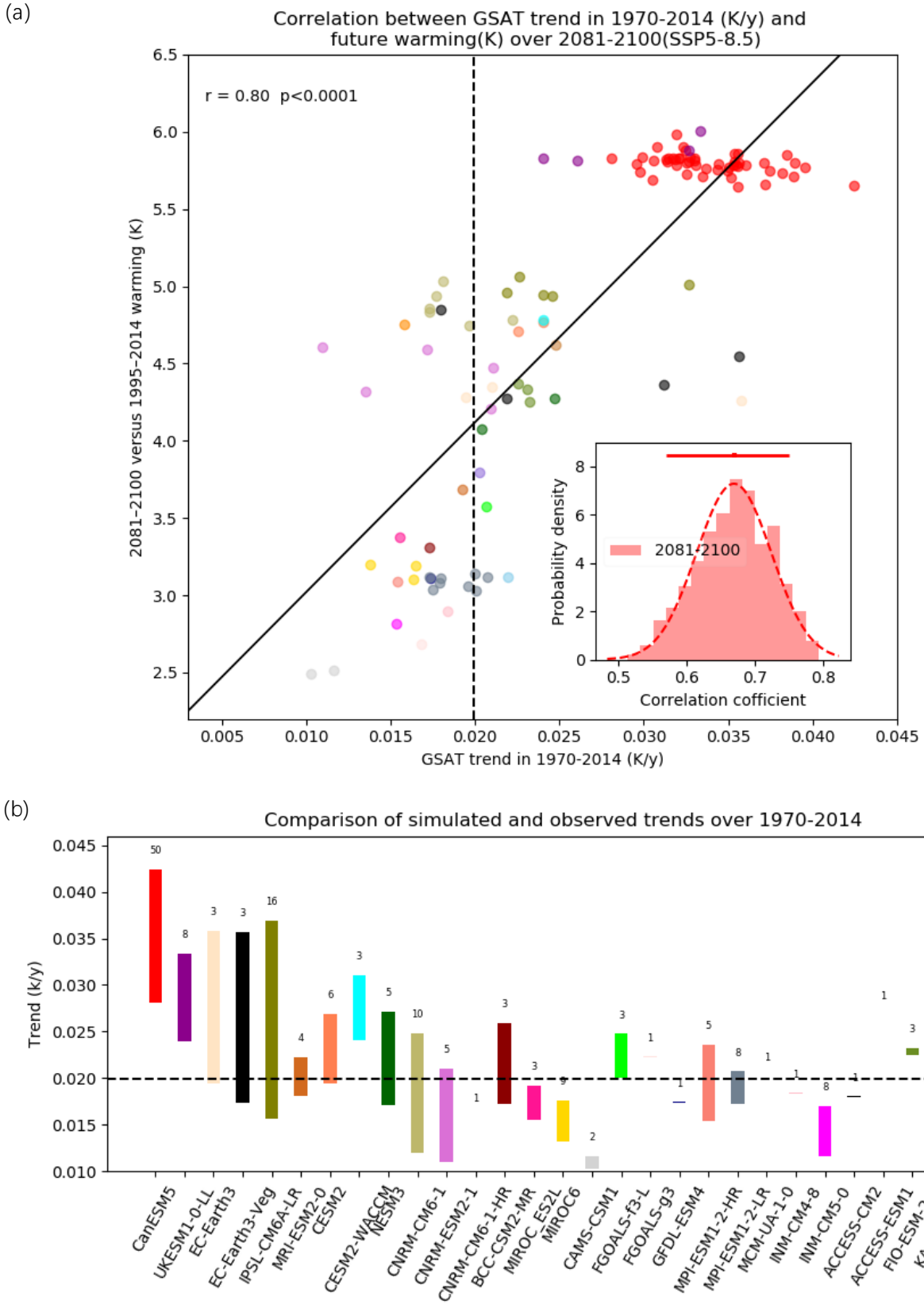


Figure 2. 1 (a) Scatterplot of projected 2081–2100 warming relative to 1995-2014 under the SSP5-8.5 scenario against simulated 1970-2014 trends in GSAT. Colors correspond to those used in panel b. Observed GSAT trends as in black dashed line. Inset:

probability density function (PDF) for correlation cofficient between GSAT trend and future warming based on 5000 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 (coloured 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 members in each model.

We also evaluate the correlation coefficient between the GSAT trend and future warming, randomly sampling one ensemble member per model (and repeating this process 5000 times; Text S4 in Appendix 1). The corner plot of Figure 2.1a 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 2.1b 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.

2.3.2 Evaluation of the weighting method

2.3.2.1 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 in Appendix 1). 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 (Fig 2.2). This procedure is repeated for 5000 random samples from the full set of model simulations. As shown in Figure 2.2a-b, 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 5000 samples is 0.004 K/y for the historical trend, and 0.04 K and 0.12K respectively 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 2.2c). 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 2.2 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 (Fig AA.S2, Table AA.S4). This result is strongly influenced by the small number of models with high climate sensitivity and large ensembles.

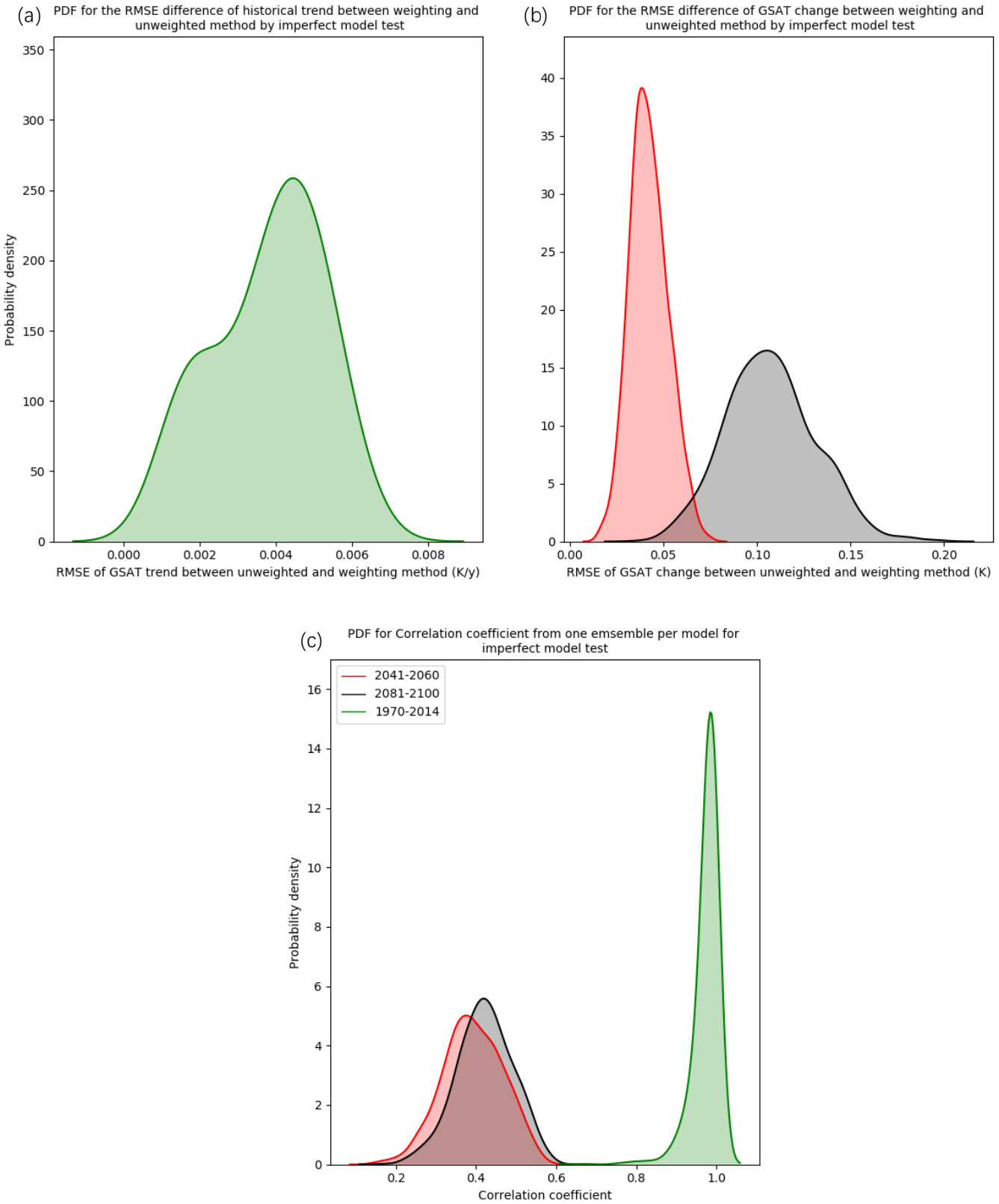


Figure 2. 2 Reductions in RMSE due to the application of the weighting approach, and correlations between mean weighted projections and pseudo-observations based on an imperfect model test with one ensemble member randomly selected per model (repeated

5000 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 ($P < 0.01$ for all 5000 samples) for the historical period, 0.40 (94% of 5000 samples show $P < 0.1$) for 2041-2060, and 0.42 (98% of 5000 samples show $P < 0.1$) for 2081-2100. The correlation coefficients of unweighted predicted means versus pseudo observation for all periods are always close to -1.

Repeating these calculations using the RMSD of gridded SAT (Text S6 in Appendix 1), 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.

2.3.3 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 in Appendix 1), 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 5000 times.

Figure AA.S4a and Figure AA.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 were 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 in Appendix 1). Qualtitatively similar results were obtained (Fig AA.S5).

2.3.4 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 Eqn (2.1) can be applied to climate change projections for which we do not have observational estimates. Figure 2.3a-d 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 2.3e). 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 5000 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 a mean of 3.82 K (3.90K) 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 2.3e, or values in Table AA.S6), are substantially lower than the corresponding unweighted values (upper bound of black shaded band in Figure 2.3e, 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 (Figure 2.3c-d); the widths of the distributions are reduced substantially by weighting.

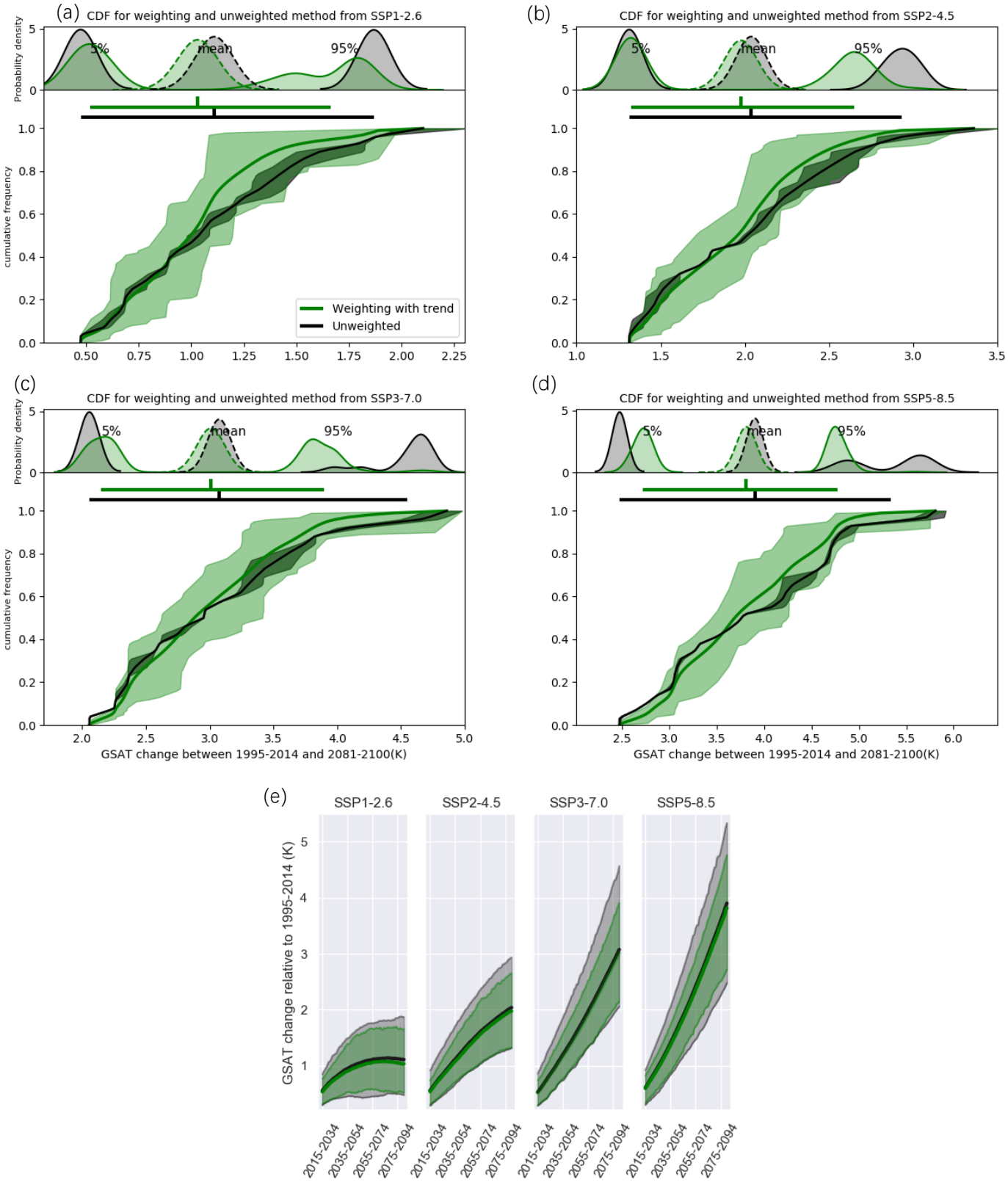


Figure 2. 3 Distributions of projected GSAT warming between 1995-2014 and 2081-2100 in each of four scenarios (Panel a-d), both constrained by observations (green) and unconstrained (black), based on 5000 samples each with one randomly selected ensemble member per model. Unconstrained projections (black) 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 (grey) results for other projection periods. The green (black) tick marks show the corresponding means of weighted (unweighted) results.

We also constrain projections using all ensemble members by calculating weights by Eqn (2.1) based on the ensemble mean for each model (Text S2 in Appendix 1) 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 AA.S6 , Figure AA.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 AA.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 AA.S7).

2.4 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 (Flynn and Mauritsen. 2020), 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 and 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 which have a realistic historical simulation also have realistic future projections (Brunner et al. 2019a; Knutti et al. 2017; Lorenz et al. 2018; Lukas et al. 2019), 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 5000 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.

Chapter 3. Emergent Constraints on CMIP6 Climate Warming Projections: Contrasting Cloud- and Surface Temperature-Based Constraints

This chapter has been published as:

Liang, Y., Gillett, N. P., & Monahan, A. H. (2022). Emergent Constraints on CMIP6 Climate Warming Projections: Contrasting Cloud-and Surface Temperature–Based Constraints. Journal of Climate, 35(6), 1809-1824.

3.1 Introduction and motivation

While multi-model mean global warming projections for comparable scenarios are similar for CMIP6 and CMIP5, some climate models in the current-generation CMIP6 ensemble project large 21st century warming and exhibit high climate sensitivity (Gettelman et al. 2019; Sellar et al. 2019; Swart et al. 2019; Voldoire et al. 2019), which is outside of the range of comparable CMIP5 projections (Meehl et al. 2020). Due to the emergence of more and higher climate sensitivity models, the CMIP6 multi-model ensemble simulates a larger spread of projected warming than CMIP5 (Nijsse et al. 2020; Tokarska et al. 2020). High levels of projected 21st century warming in CMIP6 models are associated with high equilibrium climate sensitivity (ECS), although there is some scatter in the relationship (Tokarska et al., 2020). Model differences in simulated changes of shortwave reflection by low-level clouds (particularly in the tropics and midlatitudes) in response to climate change have been found to dominate the uncertainties in global warming projections in past intercomparisons (Brient and Schneider 2016; Vial et al. 2013). For example, Brient and Schneider (2016) point out that most of the variance of CMIP5 models' ECS can be explained by model differences in low cloud-induced shortwave reflection over tropical oceans. Recent studies provide evidence that changes to the representation of cloud processes in CMIP6 models result in simulations in better agreement with satellite datasets in the midlatitudes than prior generations of models (Myers et al. 2021). However, there is a broad spread in cloud feedbacks in CMIP6 resulting in a broad range of climate sensitivity (Schlund et al. 2020; Zelinka et al. 2020b).

Statistical methods combining observations and model simulations are an effective approach to constraining projected warming and climate sensitivity. For example, Sherwood et al (2020) consider multiple lines of evidence, including historical warming and the paleoclimate record as well as process understanding of feedbacks, to estimate the effective climate sensitivity using a Bayesian framework. Ribes et al (2021a) use climate models to provide a prior on the forced response and use the observational record to obtain its posterior distribution. Hattab et al (2019) use a principal component

regression method and discuss the selection of a robust set of observable predictors to estimate climate sensitivity.

The application of emergent constraints is an effective way to narrow the projected warming spread (Bretherton and Caldwell 2020; Nijsse et al. 2020; Tokarska et al. 2020). An emergent relationship between an observable quantity in the historical or present period and a quantity related to the future climate (for example, 21st century warming) can serve as the basis to constrain projections. The emergent relationship is usually motivated by physical understanding of a process driving climate feedbacks, and observational estimates must exist to distinguish models with a realistic representation of the process from those with a less realistic representation (Bretherton; Caldwell 2020; Meehl et al. 2020; Tokarska et al. 2020). For an emergent constraint to be robust, a clear physical mechanism is required: statistical analyses in Caldwell et al. (2014) show that large correlations across models between past climate variables and future projections can occur by chance. Considering 11 metrics to constrain ECS from both CMIP5 and CMIP6 simulations, Schlund et al (2020) find that most cloud metrics first identified using CMIP5 data show a weaker emergent constraint in CMIP6. Other lines of evidence further demonstrate that not all of these metrics are robust. In particular, Caldwell et al. (2018) use a feedback decomposition analysis to evaluate the performance of several emergent constraints on ECS and show that only four of nineteen cloud constraints on ECS proposed in the literature can be considered credible. Emergent constraints using certain cloud diagnostics have resulted in higher climate sensitivity than those obtained from unconstrained ensembles. Zhai et al. (2015) constrain ECS using the sensitivity of extratropical low cloud fraction to the seasonal cycle of sea surface temperature (SST), showing that the relatively high climate sensitivity models in CMIP3 and CMIP5 are more consistent with observed values of this metric than models with relatively low climate sensitivity. By combining information from several cloud-related metrics, Bretherton and Caldwell (2020) predict a larger constrained mean of ECS in CMIP5 than that of the raw ensemble.

Although higher or medium climate sensitivity models in CMIP5 and CMIP6 are generally found to be in better agreement with observational constraints using cloud related metrics, recent analyses using the historical global mean surface air temperature warming trend as a constraint find the high sensitivity models to be less consistent and suggest that they might overestimate the future warming trend (Chapter 2; Nijsse et al. 2020; Tokarska et al. 2020). The preferential weighting of low warming projections when using the recent warming trend as a constraint is at least partially result of the so-called sea surface temperature pattern effect (associated with warming in the western equatorial Pacific Ocean and cooling in the eastern equatorial Pacific Ocean since about 1980), which has been shown to have a strong effect on the observed global mean near-surface air temperature (GSAT) trend in recent decades (Andrews et al. 2018; Dong et al. 2020; Gregory et al. 2020; Zhou et al. 2016; Zhou et al. 2021). An increase in the zonal gradient of SST across the low latitudes of the Pacific Ocean has resulted in more low cloud coverage over the eastern part of the basin and a global-scale cooling effect

due to the increased reflection of incoming shortwave radiation. As this pattern effect of east-west tropical Pacific SST gradient is more likely a result of internal variability rather than a long-term warming response, this cooling pattern is not expected to persist (Forster et al. 2021; Watanabe et al. 2021). The recent Assessment Report of the Intergovernmental Panel on Climate Change assigns a medium confidence of observed changes in the pattern effect resulting from internal variability (Forster et al., 2021). For the future changes of this pattern effect, there is medium confidence that the observed strengthening of the east-west SST gradient is temporary and will transition to a weaking of the SST gradient on centennial timescales. This SST pattern is not captured well by the CMIP5 or CMIP6 ensembles (Olonscheck et al. 2020). Compared with CMIP5, Olonscheck et al (2020) found that the use of much larger initial condition ensembles of CMIP6 models does capture the observed cooling pattern, indicating that the absence of the pattern effect in the CMIP5 and CMIP6 ensembles can be partly interpreted as a sampling bias. Therefore, the relatively low observed trend potentially induced by internal variability will favor low climate sensitivity models and may result in spuriously low warming projections when applying the observed GSAT trend as a constraint.

While past studies have focused on the use of cloud metrics to constrain ECS (Bretherton and Caldwell 2020; Brient and Schneider 2016; Brient et al. 2016; Caldwell et al. 2018; Qu et al. 2014; Zhai et al. 2015), it is unclear how effectively cloud metrics constrain the transient projected climate warming which is most relevant for adaptation and mitigation planning. This fact motivates us to investigate the use of cloud metrics to constrain CMIP6 projections of 21st century warming under various SSP scenarios. A number of issues regarding the development and implementation of emergent constraints remain outstanding. Although previous studies provide evidence of clear relationships between physically-based cloud metrics and projected warming, the relative performance of cloud metrics as constraints compared with historical warming as a constraint is unclear. In particular, the relative impacts of the internal variability on the uncertainty range of constrained projections are unclear for different categories of metrics. Furthermore, as models in the CMIP6 archive share components the simulations they produce may not be statistically independent. The impact of this potential dependence needs to be addressed when assessing the performance of the metrics. Finally, different methods for applying constraints have been proposed. For example, some studies weight models based on their performance compared to observations (Brunner et al. 2019a; Brunner et al. 2020a, 2020b; Knutti et al. 2017; Lorenz et al. 2018; Sanderson et al. 2015a, 2015b; Sanderson et al. 2017) while others use linear regression based approaches (Cox et al. 2018; Hall et al. 2019; Nijsse et al. 2020; Schlund et al. 2020; Thackeray and Hall 2019; Tokarska et al. 2020). A direct comparison of these two approaches is needed.

To address these questions, in this paper we contrast the application of cloud metrics with the application of the global surface air temperature trend as emergent constraints on CMIP6 projected warming, and apply a multivariate linear regression model to make

observationally-constrained projections of 21st century warming. We first apply a step-wise approach to select the most effective linear regression model from a subset of physically-based metrics. The selected linear model is then evaluated in a cross-validated imperfect model test. We then use the linear regression model to constrain projected 21st century GSAT changes from CMIP6 simulations. Potential model dependence is taken into account during both processes of metric selection and warming projection constraint, and different approaches to applying the constraints (weighting vs. linear regression) are directly compared.

3.2 Data and methods

3.2.1 Model simulations

We use output from 26 global climate models participating in CMIP6 (Table 3.1). We calculate the various metrics considered (Section 3.2.2) using historical simulations and then use observations with the emergent constraint approach to predict GSAT changes in 2081-2100 relative to 1995-2014 under SSP5-8.5 and SSP1-2.6 scenarios (O'Neill et al. 2016; Riahi et al. 2017). For those models which contributed initial condition ensembles, we consider all ensemble members individually (we sample one random realization per model 10000 times, as outlined in Section 3.2.4) in order to assess the contribution of internal variability, and to not bias results toward models with particularly large ensembles.

Table 3. 1 CMIP6 Historical, SSP1-2.6, and SSP5-8.5 simulations used in this study. The number of ensemble members provided for each forcing senario is indicated in the second through the fourth columns.

Model name	Historical	SSP1-2.6	SSP5-8.5
ACCESS-CM2	1	1	1
ACCESS-ESM1	1	1	1
BCC-CSM2-MR	3	1	1
CAMS-CSM1-0	2	2	2
CanESM5	50	50	50
CESM2	6	1	2
CESM2-WACCM	3	1	1
CNRM-CM6-1	10	6	6
CNRM-CM6-1-HR	1	1	1
CNRM-ESM2-1	5	5	5
EC-Earth3	3	3	3
FGOALS-f3-L	3	1	1
FGOALS-g3	1	1	1
GISS-E2-1-G	1	1	1
HadGEM3-GC31-LL	4	1	1
IPSL-CM6A-LR	16	3	5
KACE-1-0-G	3	2	2
MIROC-ES2L	3	1	1

MIROC6	9	3	3
MPI-ESM1-2-HR	5	1	1
MPI-ESM1-2-LR	8	8	8
MRI-ESM2-0	4	1	1
NESM3	5	2	2
NorESM2-LM	3	1	1
NorESM2-MM	3	1	1
UKESM1-0-LL	8	5	4

3.2.2 Emergent constraint metrics considered

We consider five potential metrics for constraining GSAT projections: four cloud-related diagnostics and the historical GSAT trend. Caldwell et al (2018) considered nineteen cloud-based metrics proposed in the literature, and found that four have both clear physical connections with cloud feedbacks and significant correlations with ECS in CMIP5. Here, we consider these four cloud-related metrics as potential emergent constraints on 21st century CMIP6 warming projections. Descriptions of these cloud metrics follow.

a. Marine Boundary Layer Cloud (MBLC) metric

The sensitivity of monthly marine boundary layer cloud (MBLC) fraction to the seasonal cycle of SST between 20°and 40°latitude in the Southern and Northern Hemisphere is well correlated with ECS across CMIP3 and CMIP5 models (Zhai et al. 2015). The fraction of MBLC is defined as the low cloud coverage (below 700 hPa) in subsidence regions (indicated by monthly climatologies of 500 hPa vertical velocity) over oceans between 20°and 40°latitude in the both Hemisphere with a random-overlap assumption (Manabe; Strickler 1964; Ramanathan et al. 1983; Stephens 1984). The MBLC metric α is defined as the regression slope in

$$\langle\overline{MBLC}(mon)\rangle = \alpha\langle\overline{SST}(mon)\rangle + \beta,\ mon = 1,2\ldots,12 \qquad (3.1)$$

The overbar and angle brackets in equation (3.1) represent the monthly climatology and spatial average, respectively. This definition of the metric is similar that of Qu et al. (2014) but uses a region which is further poleward, and focuses on monthly climatologies rather than interannual variations to reduce the uncertainty from internal variability. With future climate warming, an enhanced vertical humidity gradient between the marine boundary layer (MBL) and the free troposphere in subsidence regions can lead to greater in-cloud buoyancy and stronger production of turbulent kinetic energy in the MBL. This above process results in a weaker temperature inversion in the MBL which can decrease low cloud formation (Bretherton 2015; Rieck et al. 2012; Sherwood et al. 2014; Vial et al. 2016). As described in Zhai et al. (2015), The sensitivity of MBLC fraction to seasonal SST changes is similar to its sensitivity to centennial SST changes. The different model sensitivities of MBLC fraction to SST changes on centennial timescales account for much of the intermodel variation of climate sensitivity. Furthermore, as mentioned in Section 3.1, the larger shortwave

cloud feedback in the CMIP6 ensemble relative to the CMIP5 ensemble on average can be attributed to the changes in the simulation of mid-latitude, mixed-phase clouds [especially in Southern Hemisphere (Zelinka et al. 2020a)]. These facts provide the motivation for applying the MBLC metric as an observational constraint. For the observed value of the MBLC metric, we use the value reported by Zhai et al. (2015) of -1.28±0.19 %/K (mean ±- 1 std), estimated over the period 2006-2010.

b. Brient cloud albedo (BCA) metric

The Brient cloud albedo metric measures the sensitivity of the anomalies of shortwave cloud albedo to SST changes over the tropical oceans (Brient and Schneider, 2016). The regions considered are the monthly-varying driest quartile of ocean grid cells between 30°S and 30°N, based on 500 hPa relative humidity. In climate models, the sensitivity of the tropical low cloud albedo to the underlying surface temperature in the present-day climate correlates with the strength of the shortwave tropical low cloud feedback and with future projected warming. Thus, the sensitivity of the variation of tropical low cloud reflection to changes of SST in the present climate can be used as an emergent constraint on future GSAT changes (Brient and Schneider, 2016). A significant correlation between this metric and ECS has been found in CMIP6 simulations (Schlund et al. 2020).

c. Lower tropospheric mixing index (LTMI)

Sherwood et al. (2014) developed three metrics to measure lower tropospheric mixing: Sherwood S, Sherwood D and lower tropospheric mixing index (LTMI), all of which are correlated with CMIP5 projected warming. The Sherwood S metric quantifies climatological small-scale mixing in the tropical lower free troposphere. The Sherwood D metric quantifies the large-scale mixing over the tropical lower troposphere. LTMI is defined as the sum of Sherwood D and S metrics (Sherwood et al. 2014). Climate models with stronger vertical moisture mixing in the lower troposhere tend to have a larger increase of moisture mixing with climate warming, which could decrease the bounday layer clouds because of the stronger convective drying under climate change (Sherwood et al. 2014). Sherwood et al. (2014) demonstrated that S, D and LTMI metrics are correlated with equilibrium climate sensitivity across climate models because of this link to low cloud feedbacks. Schlund et al (2020) consider all three metrics as potential constraints, and show that only LTMI is significantly correlated with ECS in both CMIP5 and CMIP6 simulations. Therefore, in our study, we take LTMI as a potential constraint.

d. Brient cloud shallowness (BCS) metric

Brient et al. (2016) introduced the BCS metric of cloud shallowness:

$$\gamma = \frac{CF_{950}}{(CF_{850} + CF_{950})} \tag{3.2}$$

which is defined in terms of the cloud fractions below 900 hPa (CF_{950}) and below 800 hPa (CF_{850}) over weakly subsiding tropical ocean regions (vertical velocity between 10 and 30 hPa day^{-1}). Following Brient et al. (2016), CF_{950} is obtained by the mass-

weighted cloud fractions between 1000 and 900 hPa (between 900 and 800 hPa for CF_{850}).

Models that have shallower clouds over weakly subsiding tropical regions (a large BCS) in the historical period tend to have more influence by convective drying of the planetary boundary layer relative to turbulent moistening under climate warming, which further decreases low cloud cover and leads to a larger positive low cloud feedback. The detailed physical mechanism is rather complicated, involving different partially cancelling effects, and is described in detail in Brient et al. (2016). Further evidence of the relation of BCS to cloud feedbacks comes from a sensitivity test considering the lateral entrainment rate of shallow convection in MPI-ESM (Mauritsen; Roeckner 2020). Mauritsen and Roeckner (2020) found that weak lateral entrainment rates lead to more stratiform clouds within the boundary layer (larger BCS), and that convection-induced drying leads to a stronger reduction of low cloud fraction under climate warming and a higher equilibrium climate sensitivity. As our observational BCS value, we use the value based on CALIPSO/GOCOP data reported by Bretherton and Caldwell. (2020): 45%±3% over 2006–2012.

e. GSAT trend (GT) metric

Past warming simulated by climate models is well-correlated with future climate warming, especially for the 1970-2014 period (Chapter 2; Brunner et al. 2020b; Jimenez-de-la-Cuesta and Mauritsen 2019; Nijsse et al. 2020; Tokarska et al. 2020). The high correlation between GT over 1970-2014 and projected warming emerges because of the small change of aerosol forcing relative to the dominant contribution from greenhouse gases over this period. In our analysis, we use HadCRUT5 (Morice et al. 2021), which is spatially infilled, to compute the observational temperature trend (0.018±0.001 K/y, estimated over the period 1970-2014. Quoted observational uncertainty is the 5-95% range across the 200-member HadCRUT5 ensemble).

3.2.3 Linear regression and step-wise metric selection

Our aim is to integrate information from all the metrics we consider to constrain projected warming based on CMIP6 models. Our primary method for doing this uses multivariate linear regression with step-wise selection (Senftleben et al. 2020), described in detail in Appendix 2.1.1. To avoid using multiple metrics describing related processes in a linear regression model, we use a step-wise regression method to select a subset of metrics. A risk of using all possible metrics in a single linear regression model is the possibility of overfitting resulting from spurious relations between historical metrics and future projections (Bracegirdle and Stephenson 2012). We apply an iterative approach combining forward selection with backward elimination to build a multiple regression model that can best project future GSAT changes (Storch; Zwiers 1999). Specifically, this approach adds and eliminates variables iteratively to the linear regression model, stopping when the explained sum of squares (ESS) does not change significantly based on an F test (with significance level $p = 0.1$). An alternative model weighting approach (Brunner et al. 2019a; Brunner et al. 2020a; Knutti et al. 2017;

Lorenz et al. 2018; Sanderson et al. 2015a, 2015b; Sanderson et al. 2017) to constraining the ensemble is presented in Appendix 2.1.2.

3.2.4 Sampling from initial condition ensembles and uncertain observational values

About half of the models participating in CMIP6 provide ensembles of multiple initial-condition realizations. While all available ensemble members could be used together to derive an observationally-constrained distribution, this approach would give us no information about the influence of intra-ensemble variability on our results. As well, if we were to follow this approach, those models with larger initial condition ensembles would have a stronger influence on our results. Therefore, when applying the linear regression model or weighting approach, we generate the constrained projections using one randomly-selected ensemble member per model. To evaluate the influence of internal variability in the historical simulations on constrained projections, we repeat this process 10,000 times. This sampling approach is applied to all aspects of this study, including the imperfect model test and probabilistic validation (both described in section 3.2.5), and observationally constrained projections. For observationally constrained projections, along with random sampling from the initial condition ensembles, we also sample the observed quantities (X_0 in Appendix 2.1 Eq 4) within their uncertainty ranges (assuming Gaussian distributions with means and standard deviations quoted above). For each random selection of ensemble members and each realization of the observations, we calculate a PDF of future GSAT change. We then average these PDFs together to calculate an overall PDF, from which we calculate a 5-95% range of future warming (Appendix 2.1.1).

The random selection of individual model realizations is important for our analysis because internal variability contributes considerable uncertainty to observational constraints, especially for metrics based on trends (Chapter 2). The CMIP6 archive provides much larger initial condition ensembles than CMIP5, allowing us to better estimate the contribution of internal variability to projected uncertainty. Note that this sampling process can only partly account for internal variability because approximately half the models used only have a single ensemble member. To estimate the effect of undersampling internal variability, we artificially reduce the number of models with multiple ensemble members in observational constraints to test whether projection statistics are convergent as more multiple ensemble models are included. We find that for our main analysis using the full CMIP6 ensemble, the effects of internal variability may be underestimated in observational constraints using the GSAT trend metric, but this is not the case when applying climatologically-based cloud metrics (Text S3 and Figure AB.S5).

3.2.5 Imperfect model test

We apply a cross-validated imperfect model test to assess the performance of the emergent constraint approaches, comparing the linear regression and weighting approaches, as well as the metrics used in these approaches (Brunner et al. 2020b; Chapter 2). We first choose one model to act as 'pseudo-observations' (the 'truth' in the

imperfect model test context), and then apply emergent constraint approaches with all remaining models to predict this 'truth'. This procedure is repeated taking each model in turn as 'truth'. We use probabilistic validation of the imperfect model test to assess the uncertainty estimates resulting from the emergent constraints. Across all models, we note the relative frequency with which pseudo-observations lie in each quintile of the imperfect model constrained projection (0–20%, 20–40%, etc.). If the constraining approach provides well-calibrated uncertainty estimates, the relative frequency within each quintile should be close to 0.2 (Chapter 2). We also calculate the coverage frequency, defined as the percentage of pseudo-observations lying in the constrained 5-95% predicted uncertainty range. If the uncertainty estimates work well, close to 90% of pseudo-observations should fall in the 5-95% predicted uncertainty range.

There is an implicit assumption underlying our approach that cloud feedbacks not directly constrained in our study are not systematically biased in CMIP6 models. As will be demonstrated in Section 3.3.4, the agreement of constrained historical GSAT temperature evolution with observations provides evidence that the effects of any such systematic bias are limited.

3.3 Results

3.3.1 Metric performance

Based on the ECS results of previous studies discussed in Section 3.2.2, we take the BCS, BCA, LTMI and MBLC metrics as our potential cloud-related constraints. Since the GSAT trend metric has been widely considered as an observational constraint on projections, we also compare constrained projections obtained solely from the cloud constraints with those using both the cloud constraints and the GT. We now evaluate how well these observable metrics are correlated with 21st century warming using CMIP6 simulations. As shown in Fig 3.1, projected warming is significantly correlated with BCA, MBLC, and GT, such that the p-values are smaller than 0.05 across the internal variability sampling distribution (Section 3.2.4). In contrast, the LTMI and BCS correlations are weaker, with p-values larger than $p = 0.05$. These results are similar to those of Schlund et al (Schlund et al. 2020), who found that LTMI does not provide a strong emergent constraint for ECS in CMIP6 (in contrast to CMIP5). The spread of correlation coefficients is narrower for BCS, LTMI and MBLC than BCA or GT metrics. The larger spread in BCA and GT reflects a larger contribution of internal variability in contrast to the other metrics.

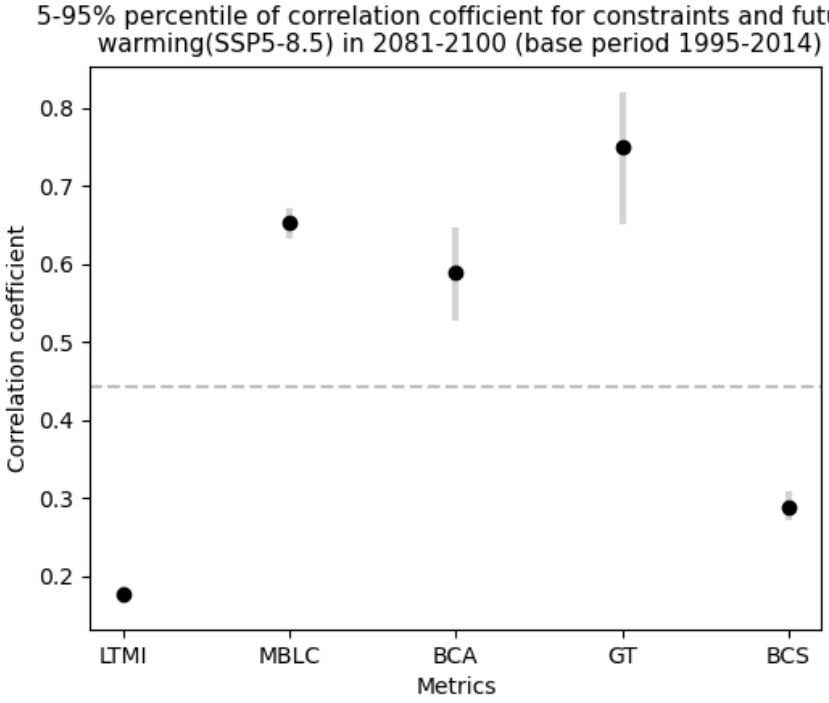


Figure 3. 1 Correlation coefficients between potential observational constraints and projected warming. 5-95% uncertainty range and mean of correlation coefficients between potential constraints, evaluated from historical simulations, and simulated warming in response to the SSP5-8.5 scenario in 2081-2100 (relative to the reference period 1995-2014) based on 10,000 random samples from the initial condition ensembles (Section 2.4). The *p*-values of the mean correlation coefficients for BCA, GT, MBLC, BCS and LTMI are 0.004, 0.0007, 0.005, 0.13 and 0.32, respectively. The horizontal grey line represents the correlation value that is significant at the 0.05 level with the number of degrees of freedom estimated based on the number of independent models (Appendix 1.3). For display purposes, the signs of the MBLC metric and BCA correlations have been reversed.

3.3.2 Step-wise regression

We now investigate which combination of metrics produces the most effective emergent constraint. We first construct a regression model using only cloud metrics. The step-wise regression approach we use adds (or removes) the most (or the least) important term in the linear regression model at each step, based on the results of F tests (Fig 3.2, Fig 3.3). For each step that requires calculation of an F statistic (Fig 3 b-d), we account for the effects of model dependence by using 20 as the effective number of independent models (Appendix 2.1.3). Following the flowchart outlining the step-wise procedure (Fig 3.2), we first build a single variable linear model with the MBLC metric because of all cloud metrics it produces the largest ESS value (Fig 3.3a). In step 2, we build a two-variable linear model adding the BCS metric because this is the only cloud metric that results in a significant increase in ESS relative to the MBLC metric regression model. As shown in Fig 3.3b, the lower 5th percentile of the F statistic range for

'MBLC+BCS VS MBLC' is greater than the critical F value which indicates a significant improvement to the linear regression model using the MBLC and BCS metrics relative to the MBLC metric only. For the other two choices 'MBLC+BCA VS MBLC' and 'MBLC+LTMI VS MBLC', the upper 95th percentile of the F statistic is smaller than the critical F-value, indicating that these two combinations should not be considered further.

In step 3 of the procedure (Fig 3.3c), we carry out a backward selection step by removing the MBLC metric. Estimates of the resulting F-statistic are always significantly larger than the critical F values (irrespective of internal variability or changes in the effective number of degrees of freedom, Appendix 2.1.3), indicating that the model with both MBLC and BCS metrics has significantly larger ESS compared with the linear model including BCS only. Therefore, the MBLC and BCS metrics are retained in the linear model in step 3.

Three-variable linear models are considered in step 4. There is no significant improvement in the fit of the combination MBLC+BCS+LTMI or MBLC+BCS+BCA relative to MBLC+BCS. The upper 95th percentile of the F-statistic in MBLC+BCS+LTMI versus MBLC+BCS and the upper 90th percentile of the F-statistic in MBLC+BCS+BCA versus MBLC+BCS are smaller than the critical F-value. While the internal variability induced range of the MBLC+BCS+BCA versus MBLC+BCS F-statistic crosses the critical F-value, the median value is much smaller. Hence, all three-variable linear models fail to increase the ESS significantly compared with the model using MBLC and BCS metrics (Fig 3.3d). Results from a sensitivity test addressing potential model dependence by varying the degrees of freedom to compute the F statistic (Appendix 2.1.3, Fig AB.S3) obtain the same set of step-wise selected metrics. Therefore, we use the MBLC and BCS metrics as constraints in our cloud metric based multiple diagnostic linear regression model to predict the GSAT changes in 2081-2100 under SSP5-8.5. We also carried out the step-wise selection on SSP1-2.6 (not shown) and get the same selected metrics as for SSP5-8.5.

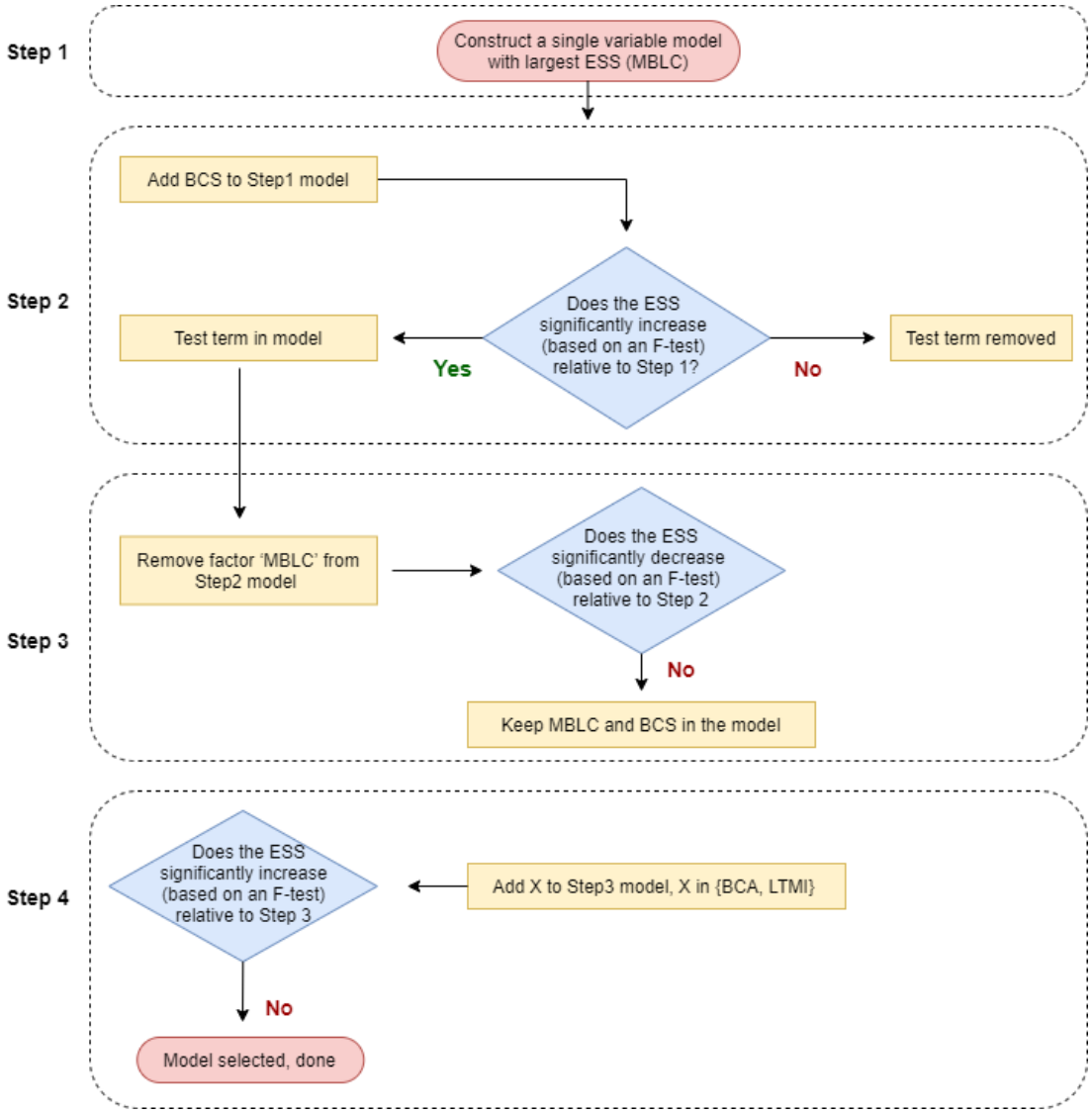


Figure 3. 2 A flow chart of the step-wise regression procedure using cloud metrics. For each step, the corresponding statistics shown in Fig 3.3.

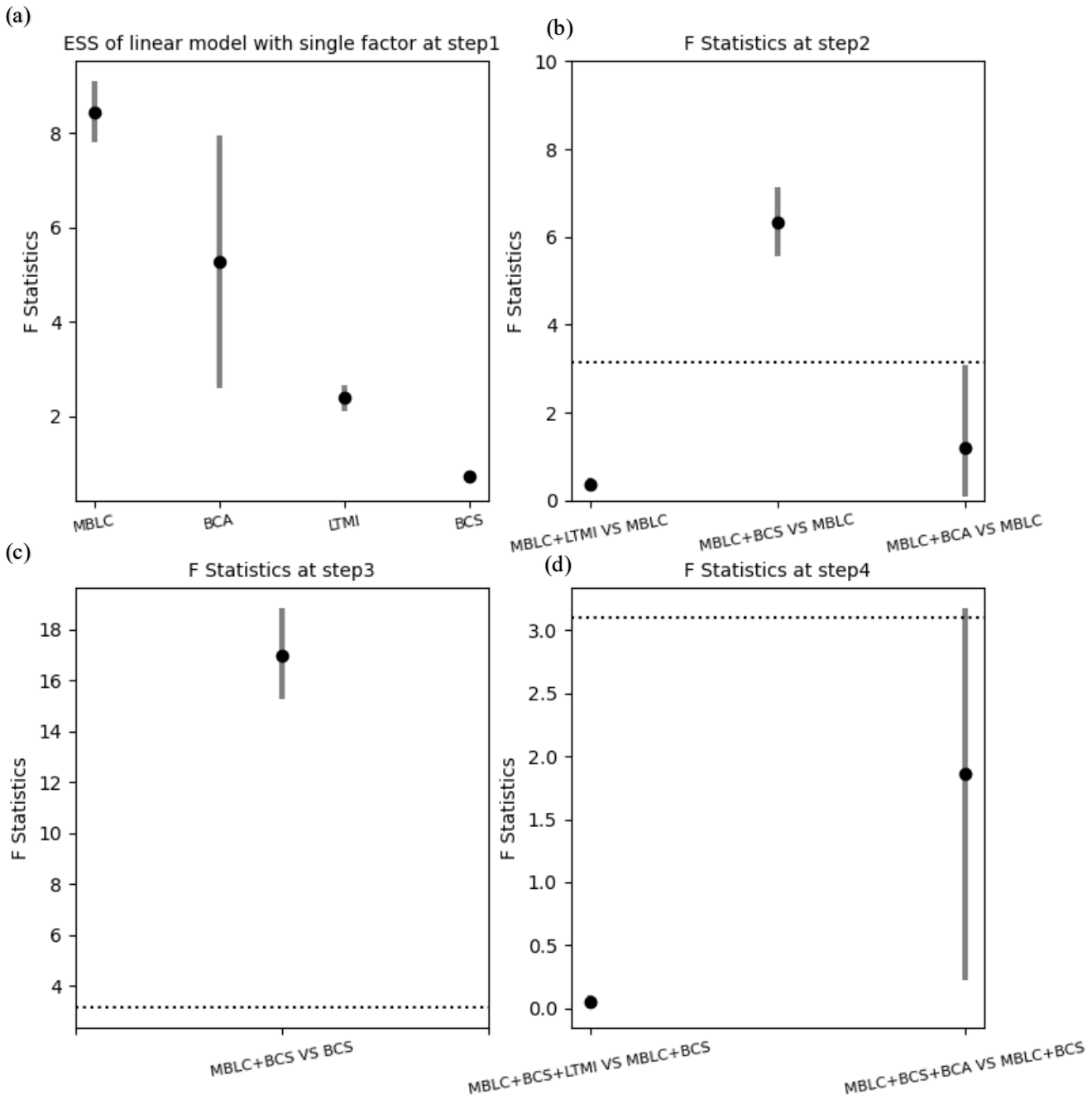


Figure 3. 3 5-95% uncertainty range and mean of F statistics at each step in the step-wise regression including only cloud metrics. The horizonal dotted lines represent critical F values at the 0.1 level. As discussed in Section 3.2.4, the 5-95% uncertainty ranges are generated by randomly sampling from the initial condition ensembles. For step 2 to step 4 in Fig 3.3, the vertical lines represent the F statistics obtained taking a value of 20 as number of statistical degrees of freedom, based on an estimate of the number of independent models in the CMIP6 ensemble (Appendix 2.2.1.3, Text S1).

Since the GT metric has been widely applied as an observational constraint in previous studies, we repeat the previous analysis including the GT metric in our stepwise regression (Fig AB.S4 of Appendix 2.2). The resulting regression model uses the

MBLC and GT metrics as constraints. A schematic of the cloud metrics entering our final regression models is provided in Fig 3.4. Although we do consider the use of GT, our main focus in the subsequent is on the use of the two metrics MBLC and BCS in our constraining process, since these two metrics are less influenced by internal variability than GSAT (also see section 3.3.3) and result in the linear regression model with best predictive power.

To further illustrate the statistical relationship between the metrics considered and projected GSAT changes in the CMIP6 simulations, Fig 3.5 shows scatter plots of each of the GT, MBLC and BCS metrics with late 21st century warming. These scatter plots each use only a single, randomly-determined ensemble member for each model to calculate the metric and the projected warming. Fig 3.5 also presents the observed values of the metrics (with uncertainty ranges). For the MBLC metric, models near the centre of the range of simulated GSAT changes are in best agreement with observations. These results suggest that applying the MBLC metric will not shift the centre of the distribution much relative to the unconstrained ensemble, consistent with previous studies of constraints on ECS (Bretherton and Caldwell 2020; Caldwell et al. 2014; Schlund et al. 2020;Sherwood et al. 2014). In contrast, observed values of the GT and BCS metrics are towards the low end of the range of simulated values. This fact is also in agreement with previous studies (Bretherton and Caldwell 2020; Brient et al. 2016; Nijsse et al. 2020; Tokarska et al. 2020).

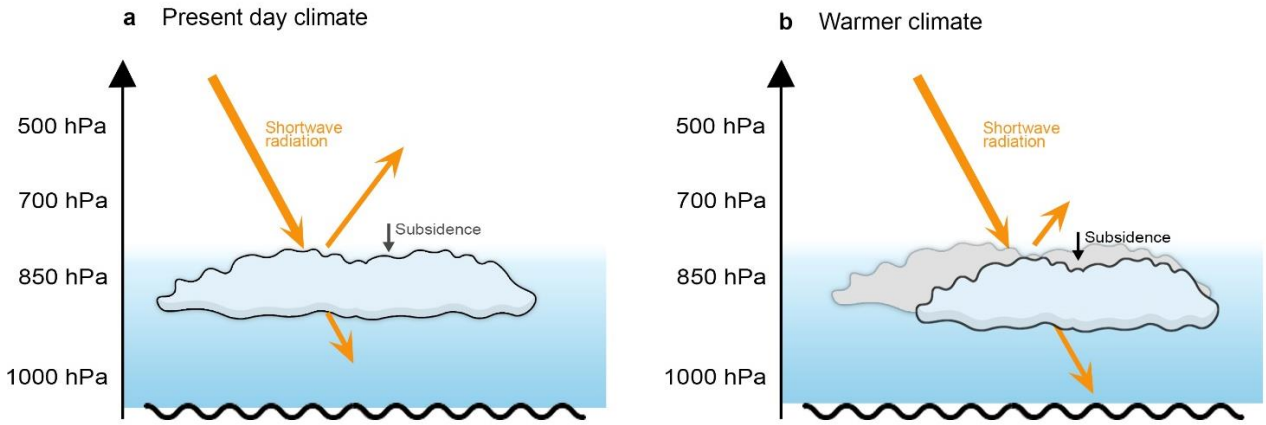
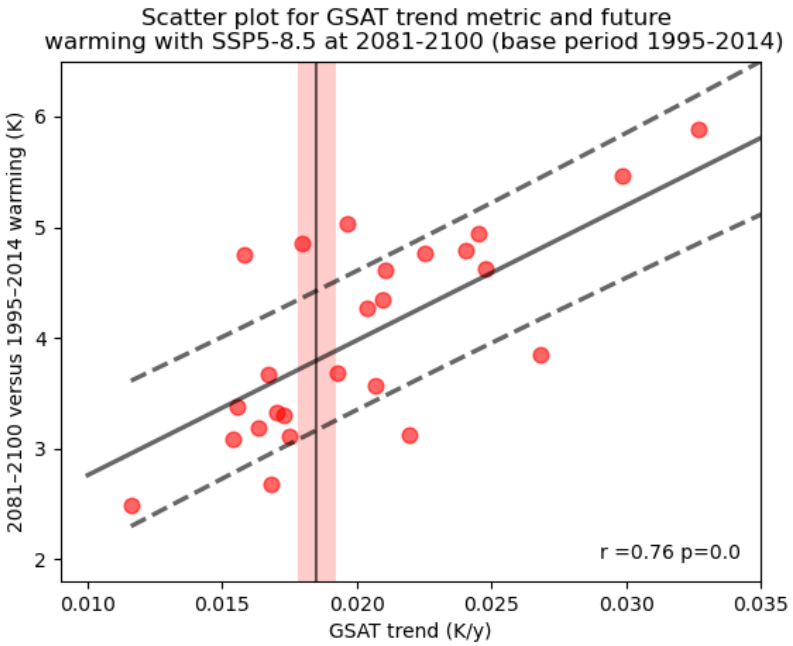
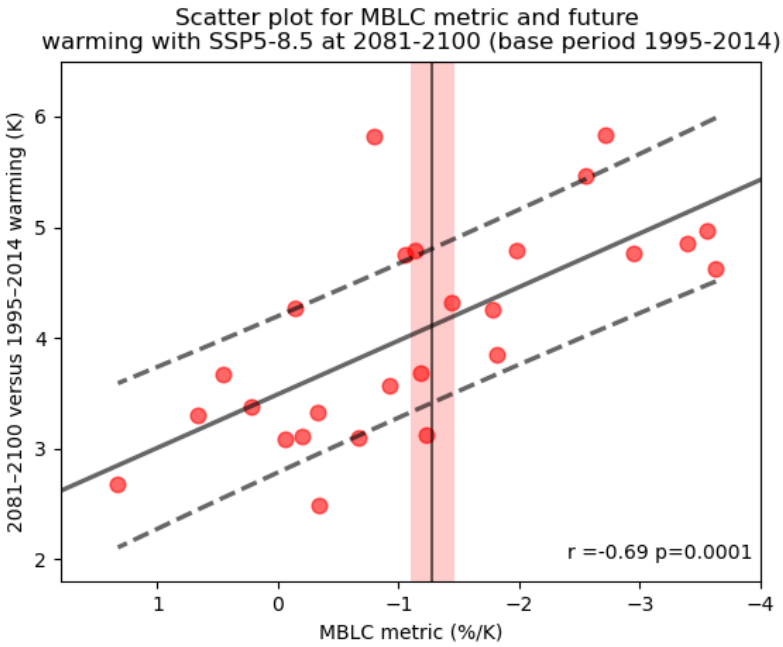


Figure 3. 4 Schematic plot showing the physical basis of BCS and MBLC metrics. For the BCS metric, models which have a stronger convective control of cloud cover in subsidence regions in the current climate tend to have shallower clouds, and tend to have a larger reduction in cloud cover associated with strengthened convective drying as the climate warms. For the MBLC metric, models with a larger decrease of MBLC fraction in response to the SST warming at seasonal scale tend to have a larger decrease of MBLC fraction to SST warming at the centennial scale. MBLC and BCS metrics focus respectively on mid-latitude and tropical low-level clouds. Both these metrics are calculated over subsidence regions over the ocean. Detailed definitions of these selected metrics are in Section 3.2.2.

(a)


Scatter plot for GSAT trend metric and future warming with SSP5-8.5 at 2081-2100 (base period 1995-2014)
2081–2100 versus 1995–2014 warming (K)
GSAT trend (K/y)
r =0.76 p=0.0
2
3
4
5
6
0.010
0.015
0.020
0.025
0.030
0.035

(b)


Scatter plot for MBLC metric and future warming with SSP5-8.5 at 2081-2100 (base period 1995-2014)
2081–2100 versus 1995–2014 warming (K)
MBLC metric (%/K)
r =-0.69 p=0.0001
2
3
4
5
6
1
0
−1
−2
−3
−4

(c)

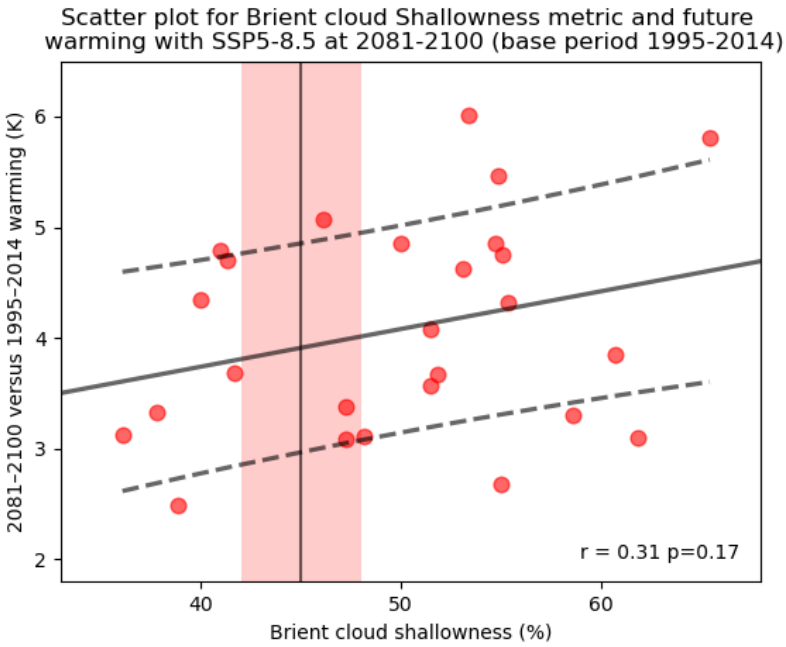

Scatter plot for Brient cloud Shallowness metric and future warming with SSP5-8.5 at 2081-2100 (base period 1995-2014)
2081–2100 versus 1995–2014 warming (K)
Brient cloud shallowness (%)
r = 0.31 p=0.17
2
3
4
5
6
40
50
60

Figure 3. 5 Scatter plots showing relationships between selected constraints and projected warming. GT, MBLC (*x*-axis reversed), and BCS metrics are respectively shown in panels a, b and c. For illustration, one ensemble member per model is used. The correlation coefficients and *p*-values (relative to a null hypothesis of no correlation) are reported in the bottom right corner of each panel. The vertical lines show the observational values with means in solid and standard deviation in shadow. The dashed lines in each panel show the 66% confidence interval of the linear regression model [Appendix 2.1 eq (5)-eq (7)].

3.3.3 Imperfect model evaluation of constrained warming

Before presenting results of projected warming constrained by observations, we evaluate the performance of the emergent constraint approach in an imperfect model setting (Section 3.2.5), based on values of the root mean square error (RMSE) improvement (relative to the unconstrained ensemble) and correlation coefficient (*r*), both calculated using the pseudo-observations and the means of the constrained imperfect model ensemble (Fig 3.6). We present linear regression results for each of the metrics separately; and linear regression and weighting results for the two pairs of metrics resulting from the stepwise selection procedure. The linear regression approach performs better than the weighting approach, resulting in larger values of both the correlation coefficient and the RMSE reduction (Fig 3.6).

Consistent with the results of the stepwise selection procedure, the linear regression models using only a single metric do not result in constraints as effective as those using two metrics (with the exception of GT, which has median correlation and RMSE reduction values similar to the combination of MBLC and BCS metrics). On average, the constrained projections based on the combination of MBLC and GSAT metrics performs slightly better than those based on the MBLC and BCS metrics. However, the uncertainty ranges (from the initial condition sampling) of both correlation coefficient and RMSE improvement are much wider for MBLC and GSAT metrics than for MBLC and BCS metrics, reflecting the substantially larger internal variability in GSAT estimates. As the slight improvement in performance of the combination of the MBLC and GSAT metrics comes with a substantial increase in uncertainty and risk of biases from using the GSAT metric due to poor representation of the pattern effect in CMIP6 models, these results suggest that the more robust constraints based on the cloud-based MBLC and BCS metrics should be preferred. Similar to SSP 5-8.5 described above, the robustness of the MBLC and BCS metrics was determined for SSP1-2.6 using an imperfect model test (not shown).

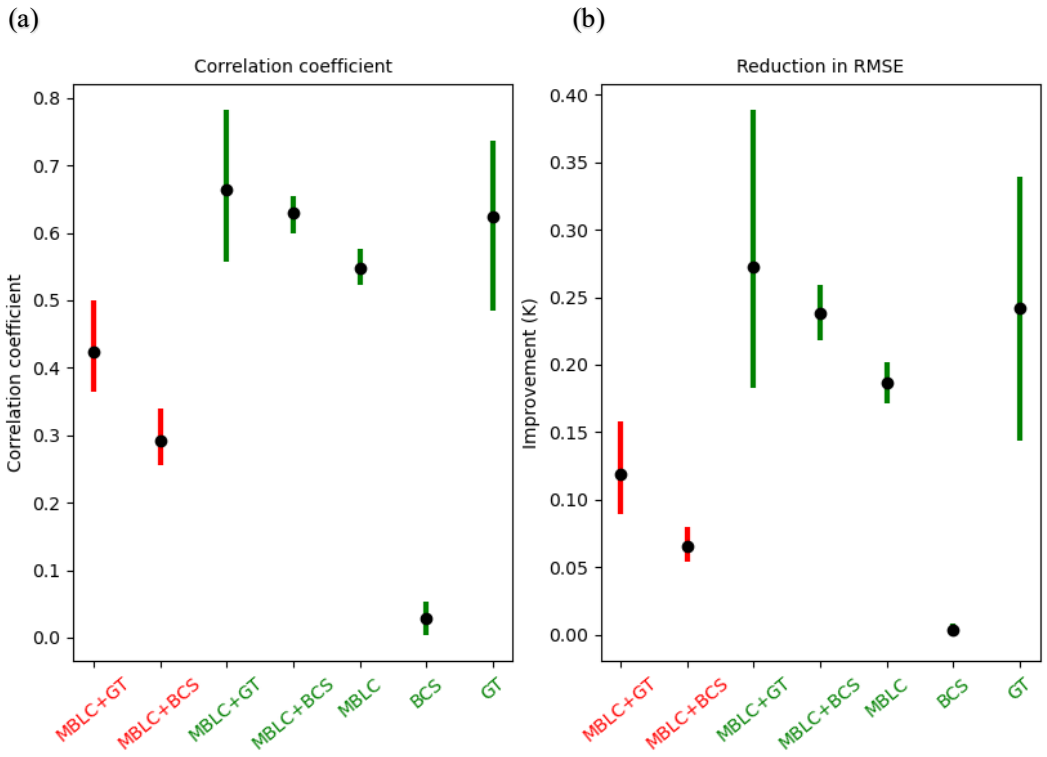


Figure 3. 6 Evaluation of constraining approaches with stepwise selected constraints using an imperfect model test. Panel (a) shows correlations between predicted means of constrained projections and pseudo-observations. Panel (b) shows reductions in RMSE of constrained projections compared to unconstrained projections. The performance of the liner regression model is shown in green bars and the performance of the weighting method is shown in red bars, using stepwise selected metrics. Please note that the 5-95% uncertainty ranges (vertical bars) and means (dots) in panel (a)-(b) are a result of initial condition sampling.

To test whether the linear regression method with the selected metrics provides reliable uncertainty estimates, we conduct a probabilistic validation (Section 3.2.5). Fig 3.7a shows the results of this analysis for 2081-2100 under SSP5-8.5 with the uncertainty estimated by assuming Gaussian regression residuals (Appendix 2.1.1.1). Also shown in Fig 3.7a are the results for the unconstrained ensemble. Similar to the unconstrained projection, the MBLC and BCS metric linear regression model produces relative frequencies of approximately 0.2 in each quintile (Fig 3.7a). As shown in Fig 3.7b, the average fraction of pseudo-observations lying within the 5-95% uncertainty range predicted by linear regression with the assumption of Gaussian residuals is close to 90%, and the range of this fraction due to sampling internal variability is narrow. These results indicate that the assumption of Gaussian regression residuals results in an accurate coverage probability.

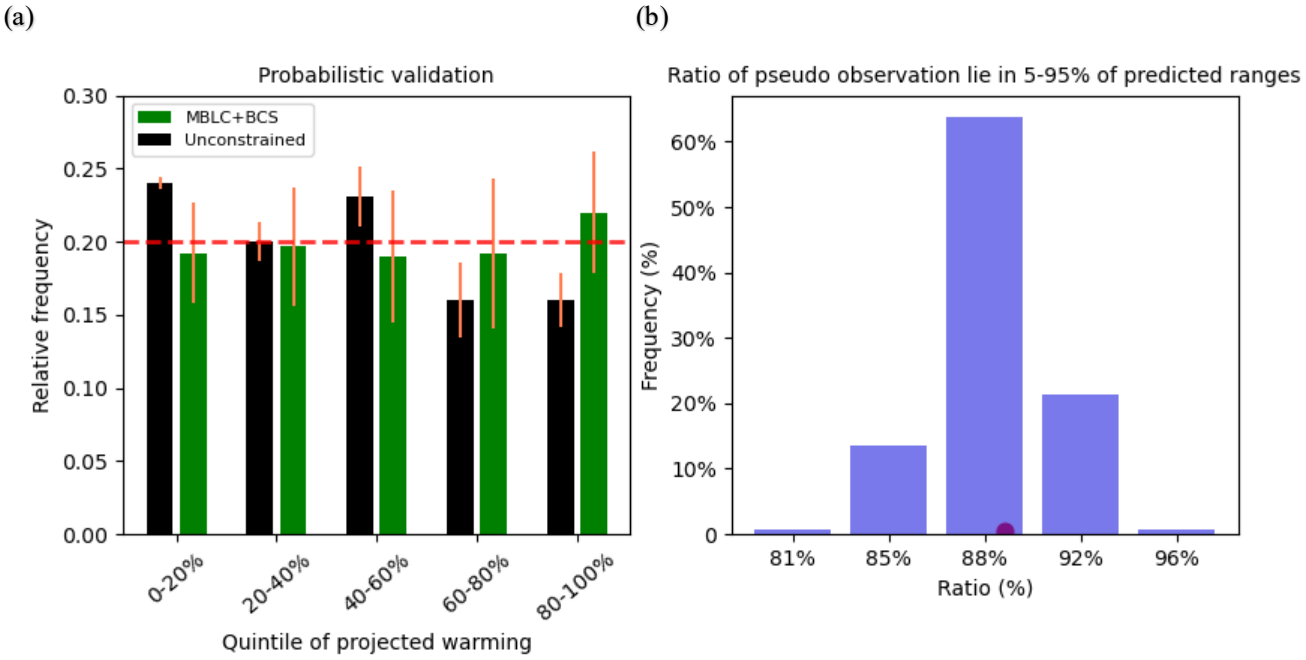


Figure 3. 7 Histograms show the relative frequency with which the true 21st century warming in the individual SSP5-8.5 simulations lies within each of five quintiles of projected warming derived using the unconstrained and MBLC+BCS constrained approaches in an imperfect model test, aggregated across all models. Bars denote the median of the 10,000 single-member per model samples. The ±1 standard deviation ranges are denoted in error bars for each quintile. Note that the constrained distributions are slightly narrower than that the unconstrained distributions. (b) The frequency of the fraction of pseudo-observations lying in the 5-95% constrained uncertainty range across 10,000 samples. The blue bars and red dot show the frequency and the mean of 10,000 samples, respectively.

3.3.4 Observational contraints

Based on the imperfect model analyses presented above, we now apply the observed metrics to constrain warming projections using the linear regression approach assuming Gaussian residuals. As described in Section 3.2.4, we account for internal variability and observational uncertainty by constructing the regression models using one randomly selected ensemble member per model (Fig 3.8) with one random realization of the observations sampled from the estimated distribution and repeating this process 10,000 times. The undersampling of internal variability resulting from the fact that only 12 models of SSP 5-8.5 have multiple ensemble members available is potentially important for GT but negligible for climatologically-based cloud metrics (Section 3.2.4). Throughout these calculations we use the value 20 as the effective number of independent models in the CMIP6 ensemble (Appendix 2.1.1.3).

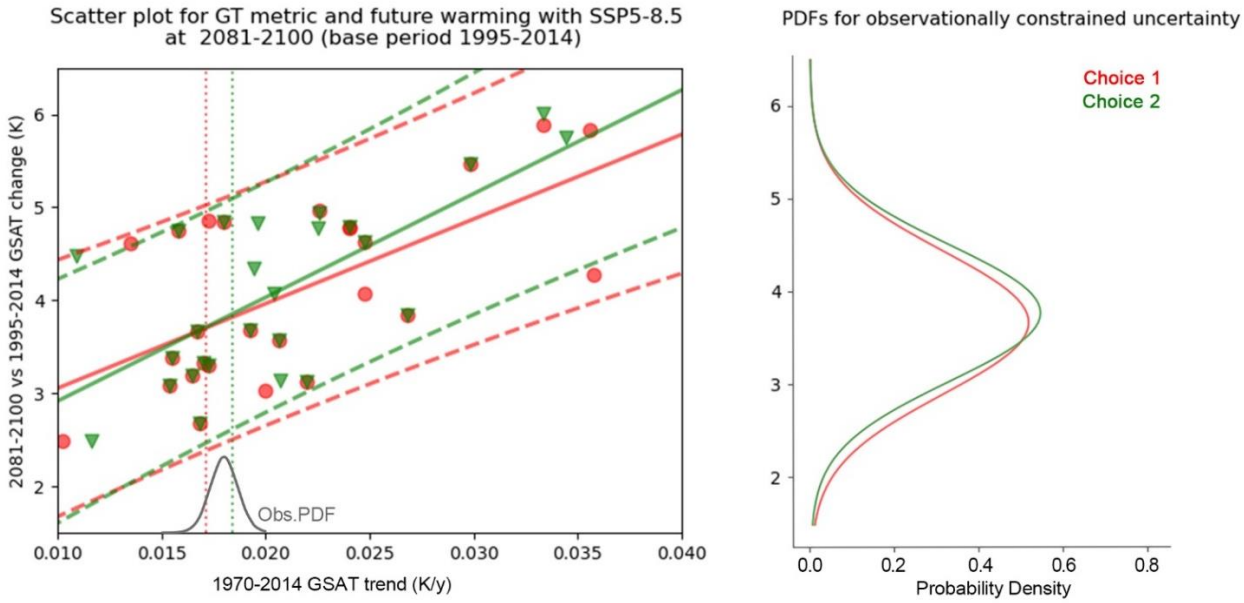


Figure 3. 8 Schematic plot to illustrate how observationally-constrained projections of warming are obtained using the GSAT trend metric with our Monte Carlo approach. The scatter plot on the left shows projected warming against historical warming in individual CMIP6 simulations, with one ensemble member chosen at random from each model. Two representative random samples of ensemble members and observations are illustrated in red and green. The associated regression relation (solid line and associated dashed lines show the linear regression model with corresponding 90% prediction interval), together with a realization of the observations (vertical dashed line), sampled from within its uncertainty range, is used to infer a PDF of projected warming, as shown in the right panel. The process is repeated 10000 times, and the corresponding PDFs are averaged to obtain the constrained projection (refer to Appendix 2.2.1.1).

Before applying the observational constraint to future climate change, we investigate how the metrics we are considering constrain historical warming in the CMIP6 models (Fig 3.9). Both the GT metric and the MBLC/BCS metric pair result in constrained historical warming consistent with observations, such that the observed GSAT always falls within the constrained 5-95% uncertainty range. This result demonstrates that the linear regression approach with the selected metrics can capture the historical warming, and increases confidence in the constrained future warming results. This is further evidence that the constrained distribution using the MBLC and BCS metrics is the absence of a systematic bias relative to observations.

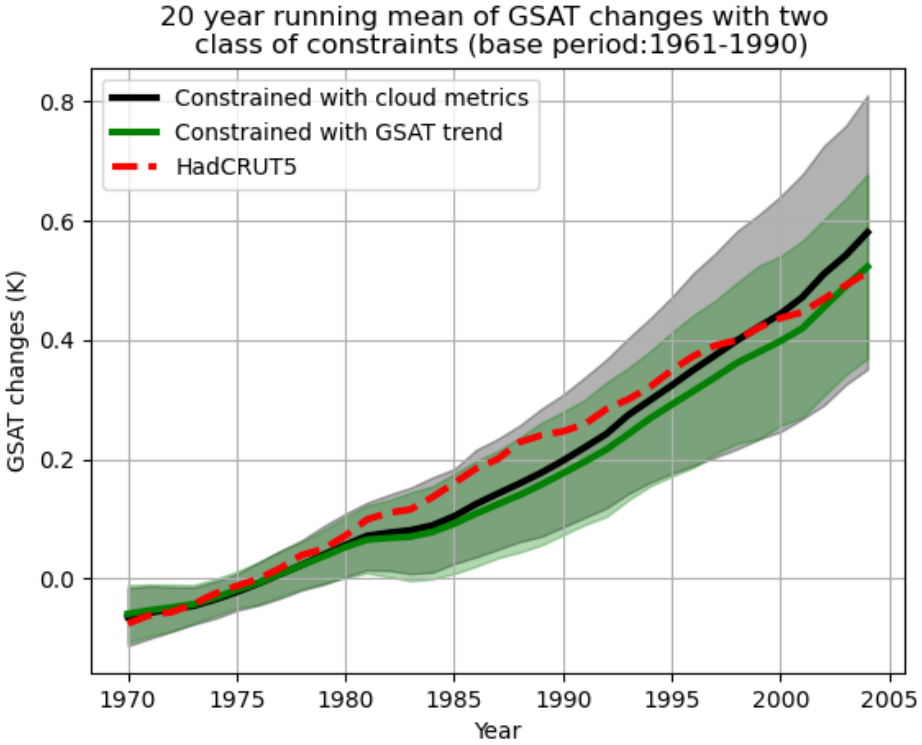


Figure 3. 9 Constrained 20-year moving average GSAT anomalies derived using the linear regression approach with each of the cloud metrics and with the GSAT trend, compared to observations (based period: 1961-1990). The observational record we use is HadCRUT5 (spatially infilled version). The x-axis shows the centre of the 20-year averaging period. The green (GSAT trend) and grey (cloud metrics: MBLC and BCS) shadows show 5-95% constrained uncertainty ranges with solid lines showing the best estimates. We account for internal variability and observational uncertainty by constructing the regression models using one randomly selected ensemble member per model and using observed quantities sampled from their uncertainty ranges (assuming Gaussian distributions with means and standard deviations quoted in section 3.2.2) and then repeating this 10,000 times.

Constraining projections of 21st century warming under SSP5-8.5 using the GT metric alone in the linear regression model produces lower values of the mean and 5th percentile of the warming distribution than either the unconstrained estimate or the constrained estimate using cloud metrics (Fig 3.10 and Table 3.2; Chapter 2; Nijsse et al 2020; Tokarska et al 2020). The 5-95% uncertainty ranges of both sets of constrained projections are narrower than that of the unconstrained ensemble.

Relative to the unconstrained ensemble, the constrained projections using the cloud metrics result in an increase in the 5th percentile of warming and a decrease in the 95th percentile, with little effect on the mean. Specifically: with the cloud metric weighting the unconstrained 5-95% uncertainty range of warming in 2081-2100 relative to 1995-2014 under SSP5-8.5 of 2.34-5.81K is narrowed to 2.84-5.12K. Assuming that the models are independent slightly decreases this range (Text S1 in Appendix 2.2, Table

3.2, and Table AB.2.S1).

Similar results are found for the projection of 21st century warming under the SSP 1-2.6 scenario (Tables AB.2.S1). Weighting by the cloud metrics reduces the 5-95% uncertainty range from 0.38-2.04 K to 0.60-1.70 K, and reduces the mean from 1.30 to 1.15 K.

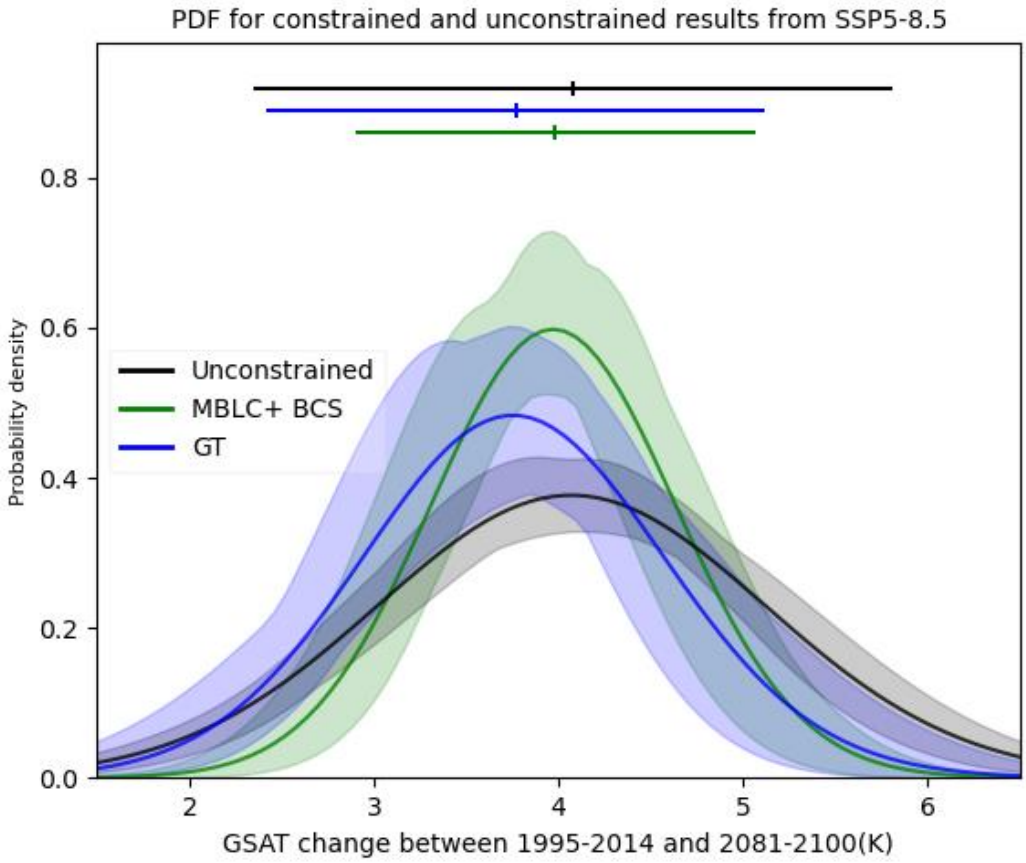


Figure 3. 10 PDFs of constrained and unconstrained GSAT changes between 2081-2100 and 1995-2014 under SSP5-8.5. The bottom panel shows the predicted distribution of GSAT changes constrained using the GSAT trend (blue), constrained using cloud metrics (green) and the unconstrained distribution (black). The shadows around these PDF curves displays the contribution of internal variability and observational uncertainty, estimated by sampling one ensemble member per model and sampling the observed quantities within their uncertainty ranges (assuming Gaussian distributions with means and standard deviations quoted in section 3.2.2) 10,000 times. The solid curves correspond to the mean of these 10,000 samples. The upper horizonal bars display the respective 5-95% projected ranges and means (numerical values are given in Table 3.2) corresponding to the mean solid curves (the theoretical basis for this calculation is shown in eq (3.8) of Appendix 2.1.1.1). These results are obtained assuming a value of 20 for the number of statistical degrees of freedom of the CMIP6 ensemble (Appendix 2.1.1.3).

Table 3. 2 Best estimates and 5-95% uncertainty ranges of projected warming using SSP5-8.5 for GSAT changes between 1995-2014 and 2081-2100. When calculating constrained uncertainty, we use a value of 20 as independent model amount in the CMIP6 ensemble (Appendix 2.1.1.3).

Metrics	Projected warming (5-95%, units: K)	
	SSP5-8.5	SSP1-2.6
Unconstrained	4.08 (2.34, 5.81)	1.30 (0.38, 2.04)
GT	3.76 (2.42, 5.11)	1.09 (0.33,1.85)
MBLC+BCS	3.99 (2.84, 5.12)	1.15(0.60, 1.70)

3.4 Summary and discussion

Previous studies have demonstrated the existence of physically meaningful relationships between equilibrium climate sensitivity and low-level cloud metrics across CMIP5 and CMIP6 models (Brient and Schneider 2016; Brient et al. 2016; Caldwell et al. 2018; Caldwell et al. 2016; Schlund et al. 2020; Zelinka et al. 2020a; Zhai et al. 2015). These relationships enable us to constrain future warming using observed values of these metrics in a complementary approach to the use of the observed GSAT trend as a constraint. We have applied two cloud metrics, a marine boundary layer cloud (MBLC) metric (Zhai et al. 2015) and Brient cloud shallowness (Brient et al. 2016), obtained from a larger set of physically-motivated cloud metrics using a step-wise selection process, to constrain future warming under two different scenarios SSP 1-2.6 and SSP 5-8.5. In developing the emergent constraints, we have compared two different approaches: linear regression (Cox et al. 2018; Karpechko et al. 2013; Nijsse et al. 2020; Senftleben et al. 2020; Tokarska et al. 2020) and a weighting method (Brunner et al. 2019a; Brunner et al. 2020a, 2020b; Knutti et al. 2017; Lorenz et al. 2018; Sanderson et al. 2015a, 2015b; Sanderson et al. 2017). Using a cross-validated imperfect model test across available CMIP6 models, we find that for the problem considered the linear regression approach produces more effective constraints than the weighting approach.

The cloud metrics we use to constrain future warming have less uncertainty which might be due to the reduced effect of internal variability than the GSAT trend often used to constrain warming projections (Chapter 2; Nijsse et al. 2020; Tokarska et al. 2020). As a result, the robustness of constrained projections based on cloud metrics is improved compared to those based on the GSAT trend metric. Furthermore, the SST pattern effect (Andrews et al. 2018; Watanabe et al. 2021; Zhou et al. 2016; Zhou et al. 2021) is known to affect climate sensitivity but is not captured well in the CMIP6 archive, perhaps due to undersampling of internal variability (Forster et al. 2021; Olonscheck et al. 2020; Watanabe et al. 2021). These considerations support the use of

the cloud metric rather than the GSAT trend for constraining future warming.

We account for observational uncertainty and internal variability in our analysis by sampling from the estimated distribution of observational uncertainty and sampling individual members from initial condition ensembles when constructing observationally constrained projections. Applying the multiple observed cloud metrics as constraints to future GSAT changes, we find that for both SSP1-2.6 and SSP5-8.5 scenarios the projected warming uncertainty ranges are considerably narrower relative to unconstrained simulations, with little change in mean warming (Fig 3.10, Fig AB.S7). We also find that observationally constrained projections using climatological cloud metrics have substantially reduced prediction uncertainty associated with internal variability in historical simulations relative to constrained projections using the GSAT trend. Furthermore, our study provides evidence for increasing the lower bound of the warming range of CMIP6 projections, as well as lowering the upper bound. This result differs from constrained projections based on the GSAT trend alone, which exhibit a substantial decrease in the upper bound and the mean of the projection range, but little change in the lower bound (Chapter 2; Brunner et al. 2020b; Caldwell et al. 2018; Nijsse et al. 2020; Tokarska et al. 2020).

Our study provides a framework to apply multiple metrics to constrain future warming which is also appropriate for constrained projections of equilibrium climate sensitivity. Our results imply that the mean climate sensitivity of the CMIP6 ensemble may not in fact be biased high as some studies have suggested, and that uncertainties in projected warming can be considerably narrowed using physically reasonable cloud constraints.

Chapter 4. Observationally-constrained projections of 21st century regional warming in the extratropical Northern Hemisphere

This chapter has been reviewed as:

Liang, Y., Gillett, N. P., & Monahan, A. H. (2023a). Observationally-constrained projections of 21st century regional warming in the extratropical Northern Hemisphere. (In review)

4.1 Introduction and motivation

Projected climate warming locally or regionally is more useful for informing adaptation planning than projected global mean warming. In the Intergovernmental Panel on Climate Change Sixth Assessment report (IPCC AR6) projections of global mean warming, ocean temperature and sea level were derived using observational constraints, rather than being directly taken from the raw ensemble of climate models. However, the report assessed that for other quantities, including projections of regional warming, 'such robust methods do not yet exist to constrain the projections' (Lee et al. 2021).

Several approaches have been applied to constrain regional climate projections. These include: weighting each model's projection based on the model's performance reproducing observations (Chapter 2&3; Brunner et al. 2019a; Brunner et al. 2020c; Knutti et al. 2017; Lorenz et al. 2018; Merrifield et al. 2020; Sanderson et al. 2015a; Sanderson et al. 2017); rescaling model projections based on a scaling factor that gives the best fit between simulated historical climate change and observations, as derived from detection and attribution methods (Allen et al. 2000; Bindoff et al. 2014; Gillett et al. 2021; Stott; Kettleborough 2002; Stott et al. 2006); Bayesian methods that update a prior distribution in light of new information provided by observations (Renoult et al. 2020; Ribes et al. 2021a; Ribes et al. 2022; Rougier et al. 2013); and linear regression approaches (Chapter 3; Nijsse et al. 2020; Tokarska et al. 2020). The observational constraints should be based on robust and physically-based connections between observable constraints in historical simulations and future projected climate change across multi-model ensembles. Such approaches have been found to be an effective way to constrain uncertainty in projections of global as well as regional warming Chapter 2&3; Thackeray; Hall 2019; Tokarska et al. 2020; Williamson and Sansom 2019)

Two types of constraint metrics have most commonly been applied in previous studies of constrained regional projections: regional metrics based on the climate of the region whose warming is being projected (Brunner et al. 2019b; Brunner et al. 2020c; Knutti et al. 2017; Lorenz et al. 2018; Ribes et al. 2022; Senftleben et al. 2020), and global-scale metrics (Hu et al. 2021; Ribes et al. 2022). The application of regional metrics to narrow the spread of future projections is based on the assumption that any over- or

under-estimation of the mean, trend or variability of the current climate in a region is most closely related to measures of projected future climate change in that region (Lorenz et al. 2018; Thackeray and Hall 2019). A number of regional climate metrics have been proposed as possible constraints on future warming (Brunner et al. 2019a; Brunner et al. 2020a; Brunner et al. 2020c; Lorenz et al. 2018; MacDougall et al. 2017; Senftleben et al. 2020). However, using a large number of possible regional constraints in a single statistical model can increase the risk of overfitting which results from spurious or non-physical emergent relations. Other risks of constrained projections exist. For example, the impact of internal variability could weaken the emergent relations, and structural errors in the relationships could produce overconfident results (Sanderson et al. 2021; Schlund et al. 2020). Two approaches can be considered to assess these concerns. A systematic metric selection process can avoid the redundancy of using multiple metrics describing closely-related processes (Chapter 3; Senftleben et al. 2020) and reduce the potential for overfitting. Secondly, the performance of constrained projections based on a range of constraints can be evaluated using model-as-truth approaches in a so-called imperfect model test (Brunner et al. 2019a; Brunner et al. 2020a; Brunner et al. 2020c; Chapter 2&3).

As discussed in Chapter 10 of IPCC AR6 (Doblas-Reyes et al. 2021), it has long been known that the pattern of projected warming in individual models stays approximately constant over time such that regional mean warming is approximately proportional to that of the global mean (Tebaldi and Arblaster 2014). Hence global constraints may also be used to constrain regional warming. Compared with regional constraints, the use of global constraints may reduce the effect of internal variability but may ignore the potential advantages of accounting for regional information (Ribes et al. 2022). The application of global metrics to constrain regional warming also disregards any relationship between inter-model differences in the spatial pattern of projected warming and inter-model differences in the pattern of warming or other climate variables in the past. To assess the relative performance of global and regional constraints we therefore compare prediction skills using these two types of metrics within a consistent evaluation framework.

In particular, this study evaluates the performance of global metrics alone relative to a combination of global and regional metrics as constraints on CMIP6 regional warming projections over extratropical Northern Hemisphere regions. We first apply metric selection strategies to select the most robust and effective regression model based on a set of regional and global metrics from the CMIP6 ensemble. We contrast the performance of the metrics selected by this approach with the performance of global metrics alone in a cross-validated test using simulations from both CMIP6 (which were used for metric selection) and CMIP5. Although these analyses are performed on different regions individually, our focus is on those predictors which produce robust and effective constraints across the extratropical Northern Hemisphere. Finally, we use the resulting linear regression models based on CMIP6 simulations, together with observed metrics to constrain projected 21st-century regional warming over sub-

continental land regions in the extratropical Northern Hemisphere.

4.2 Data and Methods

a. Region definitions, model simulations

In this study, we consider mid- and high-latitude Northern Hemisphere IPCC AR6 reference land regions [Fig 4.1; Iturbide et al. (2020)] and calculate area averages over these regions separately to compute regional metrics and projected regional warming. To make a consistent comparison in the evaluation process, we use a subset of models for which all the metrics are available. For the cross-validated imperfect model test evaluation (Section 4.3a-b), we use all available individual realizations (Table 4.1) to produce both raw and constrained projections of future warming. For the imperfect model test (described in detail in Section 4.2d), the projected warming is calculated for 2081-2100 relative to base period 1995-2014 (the base period for CMIP5 is extended by RCP 8.5 for 2006-2014), respectively considering the RCP 8.5 and SSP 5-8.5 scenarios for CMIP5 and CMIP6 simulations. The observationally constrained projections using CMIP6 multi model ensembles (Table AC.2.S1) are based on 2081-2100 relative to 1995-2014 [the same base period as in AR6 (Lee et al. 2021)] with SSP 5-8.5 and SSP 1-2.6 scenarios (Section 4.3c), for consistency with IPCC AR6 (Lee et al., 2021).

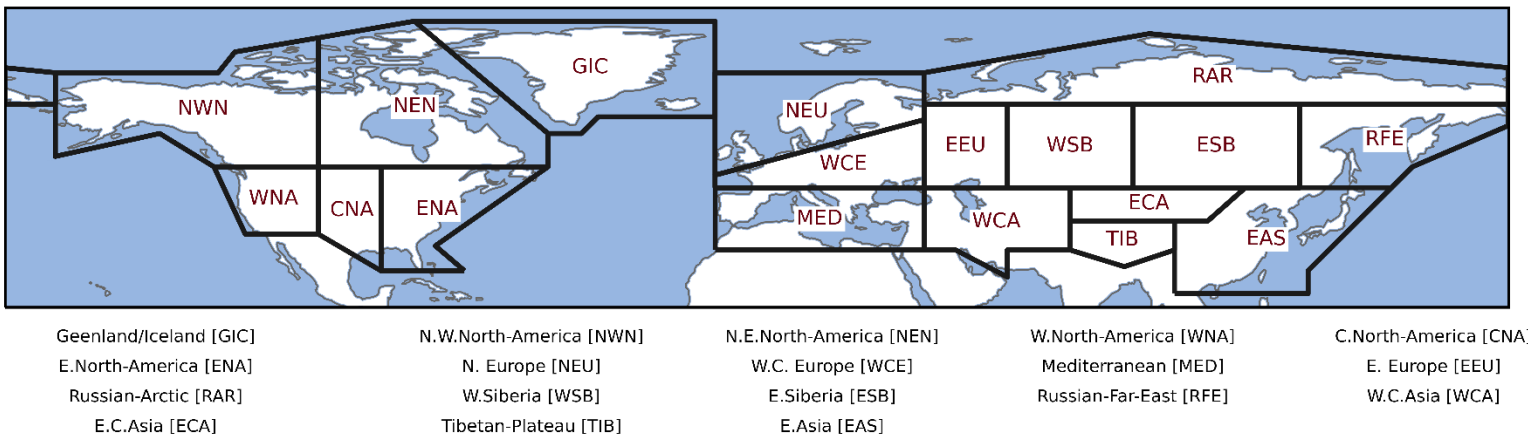


Figure 4. 1 The regions used in this study (Iturbide et al. 2020).

Table 4. 1 List of CMIP6 Historical and SSP5-8.5; and CMIP5 Historical and RCP8.5 simulations used in this paper. The numbers of ensemble members used for each experiment are listed in the second, the third, the fifth and the sixth columns. We used all simulations for which the necessary model output was available.

Model name	Historical	SSP5-8.5	Model name	Historical	RCP 8.5
ACCESS-CM2	1	1	ACCESS1-0	3	1
ACCESS-ESM1	1	1	ACCESS1-3	3	1
BCC-CSM2-MR	3	1	BCC-CSM1-1	3	1
CanESM5	50	50	CanESM2	5	5
CESM2	6	2	CCSM4	6	1

CNRM-CM6-1	10	6	CNRM-CM5	1	5
CNRM-ESM2-1	5	5	CSIRO-Mk3-6-0	1	1
FGOALS-f3-L	3	1	FGOALS-g2	1	1
FGOALS-g3	1	1	GFDL-CM3	1	4
HadGEM3-GC31-LL	4	1	HadGEM2-CC	1	1
IPSL-CM6A-LR	16	5	IPSL-CM5A-MR	3	1
KACE-1-0-G	3	2	IPSL-CM5A-LR	6	2
MIROC-ES2L	3	1	IPSL-CM5B-LR	1	1
MIROC6	9	3	MIROC-ESM	3	1
MPI-ESM1-2-HR	5	1	MPI-ESM-P	2	1
MPI-ESM1-2-LR	8	8	MRI-ESM1	1	1
MRI-ESM2-0	4	1	NorESM1-M	4	2
NorESM2-LM	3	1	NorESM1-ME	1	1
NorESM2-MM	3	1			
UKESM1-0-LL	8	4			

b. Global and regional metrics

We use a suite of global and regional metrics computed from the CMIP6 and CMIP5 multi-model ensembles (described in detail in Table 4.2). The cloud metrics are selected based on their physical connection to shortwave low-level cloud feedback which drives much of the spread in climate sensitivity between models (Zelinka et al. 2020). Several constraints based on the low-level cloud feedback have been proposed to constrain global mean projected warming (Bretherton and Caldwell 2020; Brient and Schneider 2016; Caldwell et al. 2018; Zhai et al. 2015). Chapter 3 compared the performance of regression models using combinations of the four cloud metrics identified as having the strongest physical basis by Caldwell et al. (2018) to constrain the global mean surface air temperature trend, and demonstrated that a multivariate linear regression model using two climatological cloud metrics performed best in projections of global mean warming (both in terms of cross-validated skill and robustness to internal variability). The first of these two cloud metrics is defined as the regression coefficient of monthly marine boundary layer cloud (MBLC) fraction against SST changes across the seasonal cycle over the ocean for both 20º-40º N and 20º-40º S (denoted the MBLC metric, Zhai et al. 2015). The second cloud metric characterizes cloud shallowness defined based on the ratio of cloud fraction below 900 hPa to that below 800 hPa over weakly subsiding tropical ocean regions (the BCS metric; Brient & Schneider 2016). We consider both of these global, climatologically-based cloud metrics in the present study. In addition, since the global mean near-surface air temperature (GSAT) trend over past decades is well correlated with future climate warming across climate models, we take the 1970-2014 GSAT trend as an alternative global metric. A relatively long 1970-2014 period is used for GSAT trend calculation to suppress the influence of internal variability. As aerosol forcing has not changed much over this period, the GSAT trend over this period is driven by greenhouse gas emissions, similar to future changes. Because local climate responses to changes in external forcing are found to vary approximately linearly with global mean temperature changes such that the pattern of warming remains stable

(Tebaldi and Arblaster 2014), there is a clear physical basis for constraining regional warming by applying global metrics. While the effect of the cloud metrics on regional warming will be mediated through GSAT changes, the results of Chapter 3 suggest that the climatological cloud metrics may produce more robust constraints than the GSAT trend because of the much greater internal variability of the latter.

Consistent with other regional constraint studies (Brunner et al. 2019a; Lorenz et al. 2018; Senftleben et al. 2020), we also evaluate the performance of regression models using the set of regional metrics defined in Lorenz et al. (2018) (Table 4.2). For each variable used as a regional metric, we calculate the climatology (the time mean), the linear trend, and the variability (the standard deviation), all based on the area average of the historical variable over the same regions used for the projected warming. Based on the influence of the surface energy budget on near surface air temperature (Lorenz et al. 2018), we consider several regional metrics related to radiation (surface upwelling longwave flux in air (RLUS) and surface downwelling shortwave flux in air (RSDS)), to the sensible heat flux (HFSS), and to the latent heat flux (HFLS) (Seneviratne et al. 2010). Horizontal advection of air masses and vertical motion causing adiabatic warming or cooling can influence regional warming (Meehl; Tebaldi 2004; Zhang et al. 2022). We therefore follow Lorenz et al. (2018) to use sea level pressure (PSL) as a potential predictor. Finally, we also consider precipitation (PR) as a regional constraint since earlier studies have shown that antecedent precipitation can be used as a proxy for evapotranspiration and can influence temperatures in certain regions (Hirschi and Seneviratne 2010; Mueller and Seneviratne 2012).

Table 4. 2 Short descriptions of each metric used as a constraint in this study. The global metrics (labelled as G) are MBLC, BCS and GSAT trend, while the rest of the metrics (labelled as R) are regional. The regional metrics are defined as the climatology, trend and variability of the area-averaged quantities, as described in Section 4.2b. The global (except GSAT trend) and regional metrics are calculated using data from 1980 to 2005 for both CMIP5 and CMIP6 ensembles in imperfect model test.

Name	Description	Category
MBLC	Zhai et al. (Zhai et al. 2015) defined the MBLC metric. The regression coefficient of seasonal cycle of monthly marine boundary layer cloud fraction to SST (sea surface temperature) changes over latitude band of 20° to 40° for both hemispheres. The marine boundary layer cloud fraction is obtained from the low cloud coverage (below 700 hPa) in subsidence regions (measured by monthly climatologies of 500 hPa vertical velocity)	G
BCS	Brient et al. (Brient et al. 2016) defined the BCS metric of cloud shallowness. It is defined as the ratio of cloud fractions below 900 hPa to that below 800 hPa over weakly subsiding tropical ocean regions (indicated by vertical velocity between 10 and 30 hPa day^{-1}).	G

GSAT trend	Global mean near surface air temperature trend in 1970-2014	G
HFLS	Surface upward latent heat flux	R
HFSS	Surface upward sensible heat flux	R
HUSS	Near-surface specific humidity	R
TAS	Near-surface air temperature	R
PR	Precipitation flux	R
PSL	Air pressure at sea level	R
RLUS	Surface upwelling longwave flux in air	R
RSDS	Surface downwelling shortwave flux in air	R

c. Observational constraint method and metric selection process

The main method used to constrain projections in this study is linear regression (described in Section 4.2e), which was evaluated for global mean temperature projections in Chapter 3. Two approaches are considered for selecting the metrics used for the constraint: the Lasso (least absolute shrinkage and selection operator) regression approach and the step-wise selection approach. These methods are described in Section 1 of the Appendix 3.1. For comparison with the linear regression approach, we also consider a model weighting approach (Section 2 of Appendix 3.1) proposed by Sanderson et al. (2015a, 2015b; 2017). We consider both linear regression and weighting approaches since both have been used frequently in the literature and as the different constraint methods rely on different assumptions. The weighting method assumes that models that can reproduce observations well will produce more accurate projections, while the linear regression approach is based on an emergent statistical relationship between historical metrics and projections. We evaluate and contrast the performance of these two standard observational constraint methods based on an imperfect model test (Section 4.2d). Finally, we contrast the performance of regression models based on global metrics alone with the performance of regression models based on both global and regional metrics.

d. Evaluation of the constrained projections

To assess the accuracy of the constraint approaches, we apply a cross-validated imperfect model test in which each model serves in turn as pseudo-observations which are used to constrain projections by all other models, evaluating the performance of metrics chosen based on several different measures (described below). This evaluation process has several advantages. Within the collection of CMIP models considered, we can contrast the accuracy of global metrics with the performance of constraints involving regional metrics. We are also able to assess the impact of internal variability on projections constrained using different metrics.

To test the robustness of the constraints and reduce the potential for model overfitting, we apply the metric selection process to the CMIP6 archive, and then use the selected metrics in an imperfect model test using the CMIP5 archive.

To account for the impact of internal variability, we base our constraints on one ensemble member randomly drawn from each model and repeat this random process five thousand times in the imperfect model test evaluation and observational constraint process. This approach allows us to reduce the effect of large differences in the number of realizations among different models that would be present considering ensemble means, and also allows us to estimate the impact of internal variability in the model simulations on our constrained projections by sampling across ensemble members. We also carry out a sensitivity test to assess if our constrained results are biased due to a limited number of models having ensembles of more than one member (detailed description are in Section 4.2e and Text S1 in Appendix 3.2).

In our imperfect model test, we evaluate the accuracy of the constrained mean projection using two measures. First, the correlation coefficient is calculated between constrained projections and the pseudo-observations. Because the correlation coefficient does not consider the relative magnitude of the variations, we cannot rule out that our constraining framework gives systematically biased projections even if the correlation is very close to +1. Hence, we also evaluate the reduction in RMSE defined as the difference of the root mean square error between unconstrained means and pseudo-observations from the root mean square error between constrained means and pseudo-observations. Increased accuracy of the constrained means relative to the unconstrained means is reflected in a high positive correlation coefficient as well as a positive value of the reduction of RMSE. Note that we use the unconstrained model ensemble as a baseline because this is a frequently used approach to climate projection, assuming that the real world and climate models are exchangeable.

An additional desirable property of the constrained projections is that they should have narrow uncertainty ranges, and these uncertainty ranges should be reliable. For each sample from the set of initial condition ensembles, the uncertainty range on the constrained projection is derived using the approach described in Section 4.2e) for each pseudo-observation, and the width between the 95th percentile and the 5th percentile is calculated. The mean width is then calculated across all pseudo-observations. We then compute the 5%-95% range of all constrained widths by repeating the sampling of individual ensemble members from all models 5000 times. The reliability of the constrained uncertainty ranges is measured by the ratio of pseudo-observations that lie within the constrained uncertainty ranges to the total number of pseudo-observations. The uncertainty ranges estimated are more reliable the closer the ratio of pseudo-observations that lie in the constrained 5%-95% uncertainty range is to 90%. The constrained width and coverage fraction are also used to assess the quality of different metrics chosen.

e. Observationally constrained uncertainty estimate

To construct the constrained projected warming using linear regression, we use the following steps. As in Eqn. (4.1), let the observable, constraining variable be $\boldsymbol{X}$ and $\boldsymbol{y}$ its corresponding target variable (future regional mean warming in our case). The

number of rows in $\boldsymbol{X}$ or $\boldsymbol{y}$ is the number of climate models we use. We then use ordinary least-squares regression to fit the linear model, assuming the real world and model simulations are exchangeable:

$$\boldsymbol{y} = \alpha + \boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{\beta} \tag{4.1}$$

Applying the linear regression model with observational estimates $\boldsymbol{X_0}$ of metrics chosen yields a constrained best estimate of regional projected warming $\hat{y}_0$ (Hooper and Zellner 1961; Karpechko et al. 2013; Senftleben et al. 2020),

$$\hat{y}_0 = \hat{\alpha} + \boldsymbol{X_0^T}\hat{\boldsymbol{\beta}} \tag{4.2}$$

where the scalar $\hat{\alpha}$ and the vector $\hat{\boldsymbol{\beta}}$ are the multiple regression coefficients. Assuming Gaussian residuals for predicting constrained regional warming changes (y) given $\boldsymbol{X_0}$, the probability density function (PDF; Hooper and Zellner 1961) for y is

$$p(y|\boldsymbol{X_0}) = \frac{1}{\sqrt{2\pi\sigma_{\hat{y}_0}^2}} \exp\left(-\frac{(y-\hat{y}_0)^2}{2\,\sigma_{\hat{y}_0}^2}\right) \tag{4.3}$$

where

$$\sigma_{\hat{y}_0}^2 = s^2(1 + \boldsymbol{X_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}) \tag{4.4}$$

and

$$s^2 = \frac{1}{M-p-1}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2 \tag{4.5}$$

In Eqn. (4.5), M denotes the number of models for CMIP6 archive and p is the number of metrics chosen.

Our method considers different contributions to constrained uncertainty as in Simpson et al. (Simpson et al. 2021). To account for the contribution of internal variability in constraints and the contribution of observed uncertainty to constrained uncertainty, we sample over internal variability by drawing one ensemble member from each model randomly and drawing $\boldsymbol{X_0}$ from a distribution representing observational uncertainty to estimate the regression coefficients α and $\boldsymbol{\beta}$, and repeat this process 5000 times. The regression model parameter uncertainty is accounted for in our regression model uncertainty, which corresponds to the second term in the brackets in Eqn. (4.4) above. Because our study uses one ensemble member per model [y_m in Eqn. (4.5)] rather than ensemble means, the internal variability in the future contributes to the variance in the regression residuals, as represented by Eqn. (4.5). Besides internal variability in the projections, other factors contributing to spread of constrained uncertainty, such as model uncertainty, will also contribute to the variance in the regression residuals in Eqn. (4.5). Hence the unexplained component of the variance is also accounted for in constrained uncertainty.

Each of the 5000 samples yields a PDF of constrained projections with a particular

mean (μ) and standard deviation (σ). Multiplying the joint distribution of these two statistics $f(\mu,\sigma)$ by the conditional distribution $f(y|\mu,\sigma)$ obtained from the constraints, and then integrating over μ and σ, gives us a population estimate $g(y)$ of the marginal distribution of the projected warming:

$$g(y) = \iint f(y|\mu,\sigma) f(\mu,\sigma) d_{\mu} d_{\sigma} \quad (4.6)$$

Sampling μ and σ from their joint distribution and then averaging the resulting conditional distributions of y gives a sample estimate of this population mean distribution. We then estimate constrained uncertainty based on the estimate of $g(y)$.

We also test if our constrained uncertainty estimated by our one ensemble member per model sampling strategy (described in Section 4.2d) is biased due to a limited number of models having ensembles of more than one member (Text S1 in Appendix 3.2). We conduct a sensitivity test based on two sets of synthetic data. We construct the first set of data by producing synthetic 50-member ensembles for both historical predictor and future predictand of each model by assuming Gaussian distributions for all members of the multi-model ensemble. The Gaussian distributions are centered on the ensemble means of individual models with standard deviation taken from CanESM5. The second set of data is generated in the same way as the first set but for each model the ensemble size is the same as in Table 4.2. The constrained PDFs obtained from each set of synthetic data (Section 4.2d) are then compared. The synthetic test indicates that the limited ensemble sizes available do not have much effect on our resulting PDF by using potential metrics. More detailed information is provided in supplementary Text S1 in Appendix 3.2.

4.3 Results

a. Strength of emergent relations

We first investigate the strength of the statistical relationship between projected regional warming and the global and regional metrics by computing their correlation coefficients. We evaluate the correlation coefficient between each current climate metric and projected regional warming for each region in simulations from both CMIP6 (Fig 4.2) and CMIP5 (Fig AC.S2), sampling over both models and internal variability. We sample over models by selecting (with replacement) 20 CMIP6 models (18 CMIP5 models) from the available multi-model ensemble, and we subsequently sample over internal variability by choosing one ensemble member at random for each model. Sampling over models can account for the limited number of available models and assess the increased chance of spuriously high correlation coefficients from a small sample of models. We take a 90% confidence interval on the correlation coefficient which does not intersect zero as implying a statistically significant correlation.

For CMIP6, the GSAT trend and cloud metrics MBLC and BCS generally correlate well with regional mean warming across the most regions (Fig 4.2). The correlation coefficient values for the BCS metric are not as large as GSAT trend or MBLC,

especially for the EEU, MED, RFE and WCA regions. For CMIP5, correlation coefficients with these metrics are somewhat weaker, especially for MBLC (Fig AC.S2) over GIC, TIB and EAS region, although the sign of the correlation coefficients remains the same.

By contrast, the Lorenz et al. (2018) regional metrics do not in general exhibit consistent significant correlations with projected warming across regions, especially the climatological mean and standard deviation metrics (Fig 4.2 and Fig AC.S2). While some trend-based regional metrics show statistically significant correlations in some regions (for example TAS and RLUS), there are many differences between regional correlation values in CMIP5 and CMIP6, indicating non-robustness of the statistical relationships. Our study does not sample over models in subsequent analyses, because we account for uncertainty related to the limited sample size of models as in in eq (4.4).

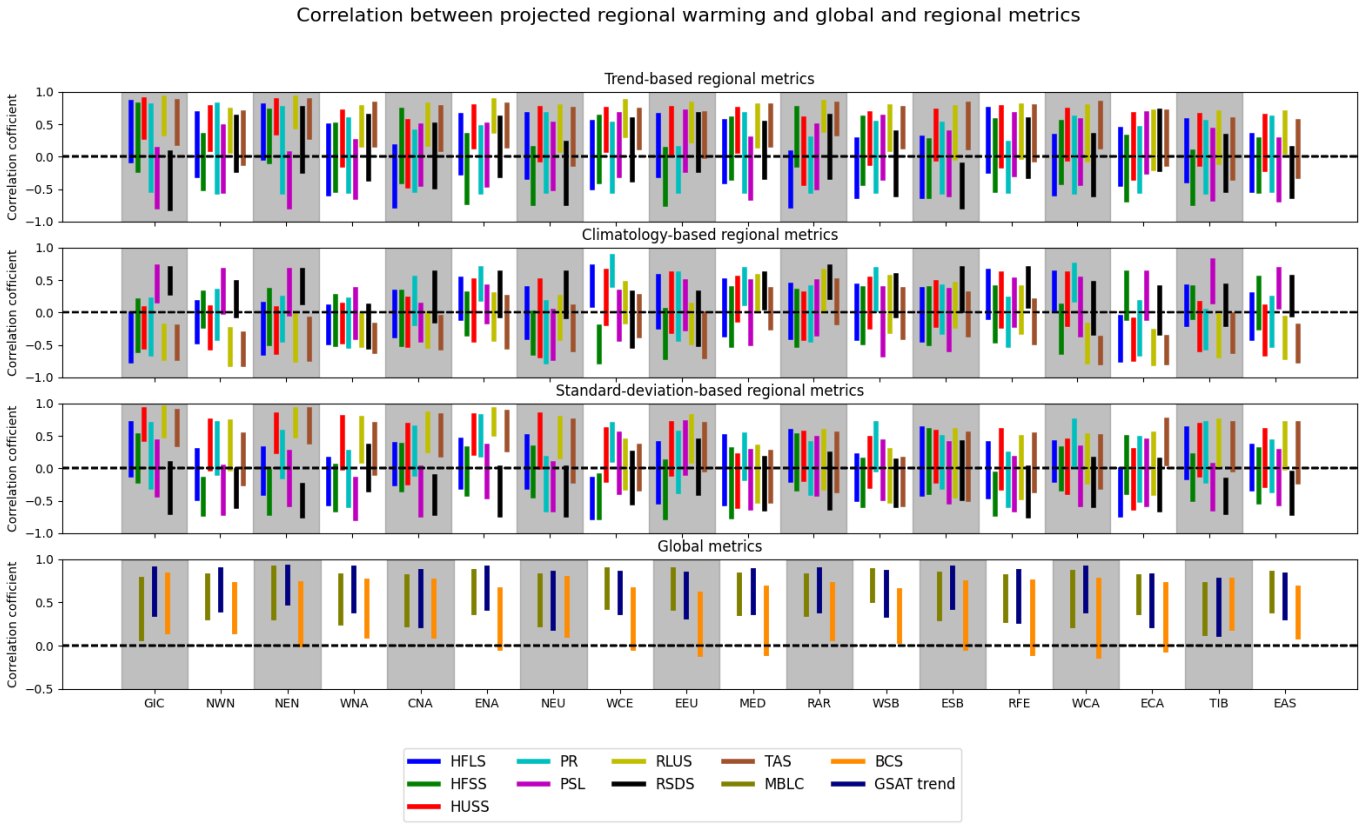


Figure 4. 2 Correlation coefficient between historical predictors and future regional warming by applying bootstrap resampling over models and initial condition ensembles (1000 times). 5-95% uncertainty ranges are shown. The sign of the correlation coefficient of MBLC is reversed. The bottom row refers to regions as defined in Figure 4.1 and metrics defined in Table 4.2

b. Imperfect model evaluation of constraints

Although our initial analysis (Figure 4.2) suggests that regional metrics are less likely to provide robust constraints on future regional warming, it is still possible that combinations of regional and/or global metrics might outperform global metrics alone.

Therefore we evaluate the performance of constrained projections produced using linear regression using metrics chosen from the set of global metrics and regional metrics (Table 4.2) and using both Lasso and stepwise-selection approaches (the selected predictors are shown in Fig AC.S3). Fig 4.3 shows the resulting constrained means as measured by the correlation coefficient and reduction in RMSE (cf. Section 4.2d). Consistent with results for global mean warming projections (Chapter 3), the linear regression method (red bars in Fig 4.3a-b) gives more accurate constrained means than the Sanderson weighting approach (blue bars in Fig 4.3a-b) across regions, both using the same global cloud-based constraints MBLC and BCS. With only one exception (TIB), the Sanderson weighting approach gives lower correlation coefficients and a lower reduction of RMSE than the linear regression approach across all regions examined (Fig 4.3 a-b). Based on this result, we use the linear regression approach for the remainder of the study.

While the performance of the Lasso and step-wise metric selection are similar for CMIP6 (Fig 4.3 c-d), when tested with models from CMIP5 (Fig 4.3 e-f), the metrics chosen with the Lasso approach generally perform better than those chosen with the step-wise selection approach across most regions, based on correlation and reduction of RMSE values. This result shows that applying step-wise metric selection rather than Lasso increases the possibility of overfitting. Therefore, we use the Lasso approach for the rest of this evaluation.

There are several clear pieces of evidence that show a lack of robustness of constrained projections using regional metrics relative to those using global metrics. First, the split sample test shows that overfitting may occur when using regional metrics. As shown in Fig 4.3e-f, for the split sample tests, the global metrics generally out-perform the regional metrics, even those selected based on Lasso. Furthermore, the split sample test shows the emergence of overfitting in more regions when using regional metrics, with several regions showing increases in RMSE. This effect also happens with global cloud metrics, but only in the TIB region, suggesting an inconsistent warming pattern across models in this region. Second, the choice of regional constraints varies depending on the metric selection approach. As shown in Fig AC.S3, different regional metrics are selected by the different methods without consistency across regions while global metrics are consistently selected. In addition, as shown in Fig 4.3c-f, the in-sample and split sample correlation coefficient and reduction in RMSE from the imperfect model test display a large spread when involving regional metrics. However, the ranges for global metrics are much narrower (especially for cloud metrics), indicating a much greater influence of internal variability using regional metrics.

The GSAT trend metrics do not outperform cloud metrics across regions, particularly in the split sample test, and the application of the GSAT trend metric displays a larger influence of internal variability. In addition, the GSAT trend metric is not chosen for most regions in the metric selection process (Fig AC.S3), while cloud metrics are chosen for most regions, indicating the cloud metrics are more robust over Northern

Hemisphere regions. For these reasons, we use cloud metrics in the rest of this study.

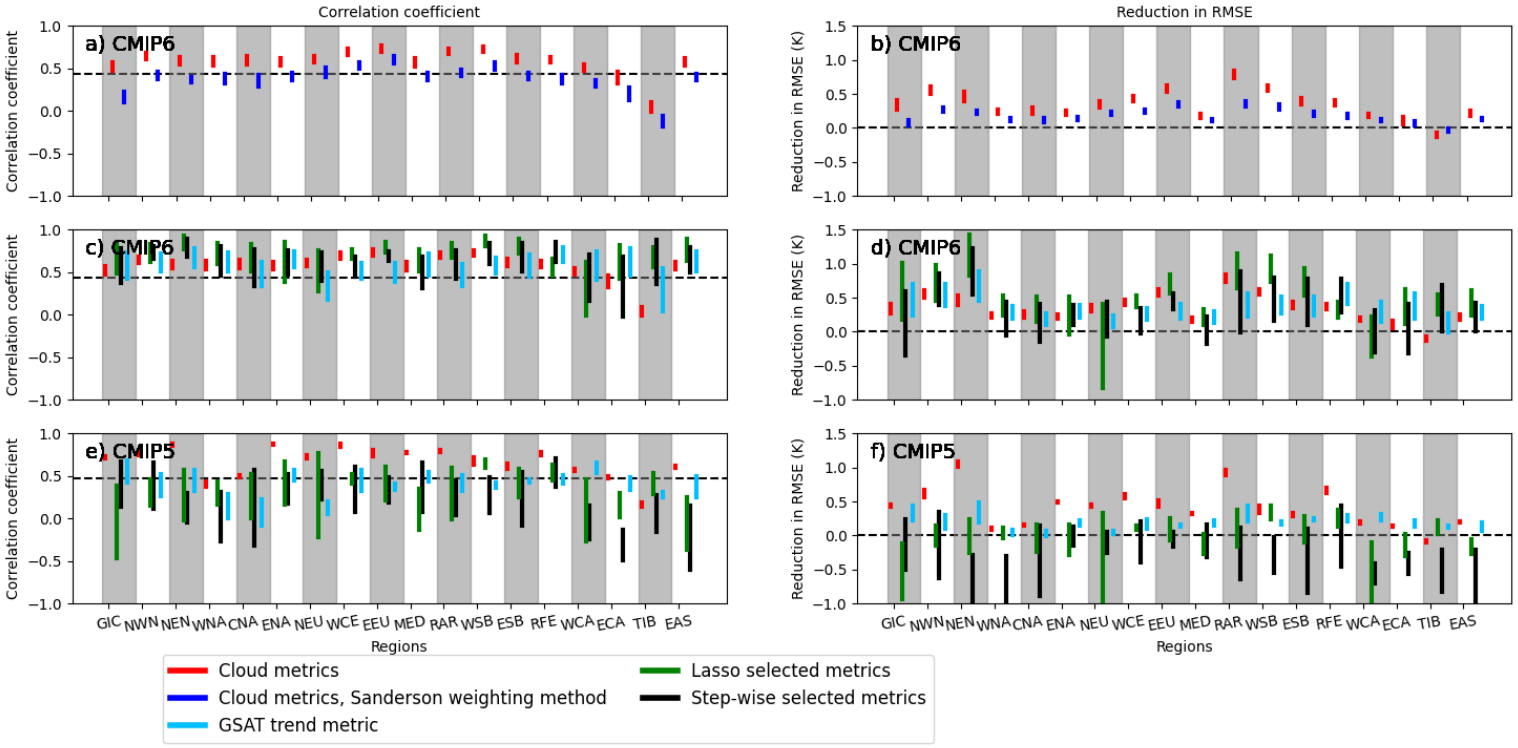


Figure 4. 3 Imperfect model test of the accuracy of the constrained temperature changes for 2081-2100 relative to 1995-2014 using SSP5-8.5 for CMIP6 and RCP 8.5 for CMIP5. The left panels show correlation coefficients calculated between pseudo-observations and projected means and the right panels show the RMSE reduction calculated as the RMSE of unconstrained means relative to pseudo-observations minus the RMSE of constrained means relative to pseudo-observations. a) and b) compare the regression and Sanderson weighting approaches using cloud metrics and CMIP6 data. c) and d) compare the performance of different sets of metrics using the regression approach and CMIP6 data, and e) and f) show the same comparison using CMIP5 data. Bars show the 5th-95th percentile range across 5000 samples, sampling across model ensemble members. The black dashed line in the left panel plots show the critical value of the correlation coefficient that is significant at the 0.05 level (one-sided) for CMIP5 and CMIP6 respectively. The black dashed line in the right panel plots shows the threshold value of 0 representing no improvement based on the metrics chosen. All color bars shown in the plot are based on the linear regression method described in Section 4.2e except the dark blue bars which are based on Sanderson weighting approach using MBLC and BCS constraints.

We further evaluate the constrained uncertainty range in projected regional temperature change estimated using cloud metrics alone or using both global and regional metrics. Using the imperfect model test, we assess whether the constraining process can reduce the uncertainty range (Fig 4.4) and evaluate if the estimation of constrained uncertainty is reliable (Fig 4.5). As shown in Fig 4.4, linear regression models including both regional and global metrics produce narrower uncertainty ranges than those based

solely on global metrics for almost all regions in the in-sample test. However, for the split sample test, uncertainties are not systematically narrower when using both regional and global metrics. Use of cloud constraints substantially and robustly reduces the projected uncertainty range relative to unconstrained projections across regions in both in-sample and split sample analyses. As shown in Fig 4.5, the constrained uncertainty estimated using cloud constraints shows a reasonable coverage ratio of around 90% (the ratio of pseudo-observations that lie in the constrained uncertainty range) across most regions. Relative to the in-sample results, metrics involving regional constraints show reduced performance of the coverage ratio in many regions in the split sample test.

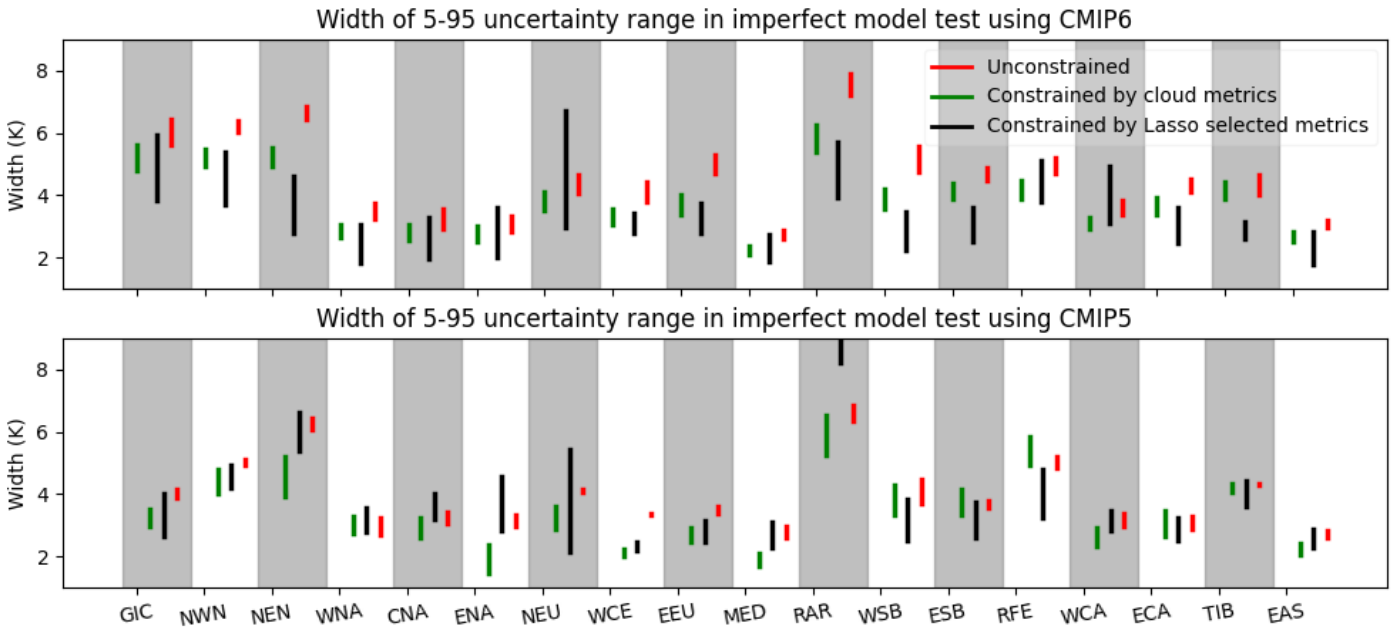


Figure 4. 4 Widths of constrained and unconstrained 5%-95% ensemble uncertainty ranges in projected regional warming. The bars show 5th-95th percentile ranges of ensemble widths based on 5000 random selections of model ensemble members.

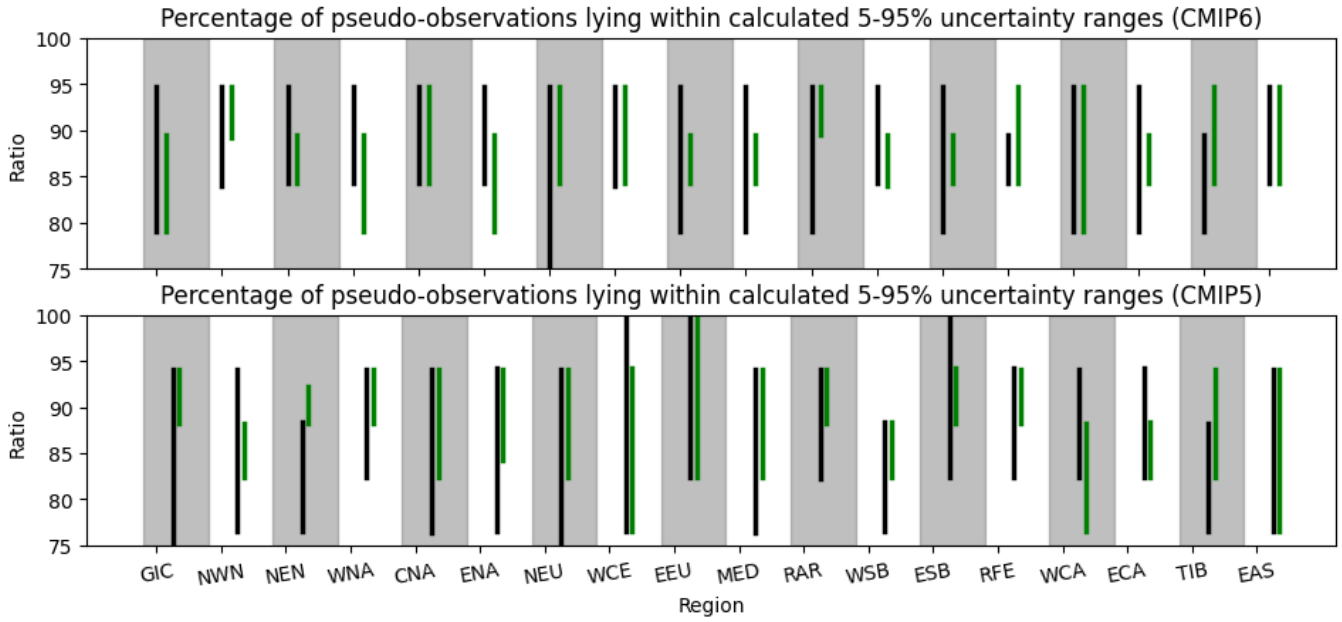


Figure 4. 5 Evaluation of the reliability of the uncertainty estimates on the constrained projections. The upper panel shows the percentage of pseudo-observations lying within constrained uncertainty ranges (or coverage ratio) for the in-sample test based on CMIP6 simulations and the lower panel shows corresponding results for the split sample test based on CMIP5 simulations. The bars show 5-95th percentile ranges based on 5000 samples of initial condition ensembles. The bars in black represent constrained projections using Lasso selected metrics and the bars in green represent constrained projections using cloud metrics.

Finally, we use the imperfect model test to evaluate the application of cloud metrics to constrain 2081-2100 warming under SSP1-2.6 (Fig AC.S4 -S6). Based on the analysis of constrained SSP 5-8.5 projections, we only uses cloud metrics. We find that with the exception of only one region (TIB), applying cloud constraints with SSP1-2.6 output results in substantial improvements relative to unconstrained projections in the mean (Fig AC.S4), and the width of the constrained projection distribution (Fig AC.S5), with reliable constrained uncertainty estimates (Fig AC.S6).

c. Observationally constrained projections

The evaluation process described in Section 4.3a-b shows that constrained projections based on cloud metrics alone are more accurate, robust, and reliable than those which also use regional metrics. They are also effective in reducing the widths of projected temperature change ranges. Following the procedure described in Section 4.2e, we accordingly use cloud metrics with observed MBLC reported by Zhai et al. (2015) of $-1.28\% \pm 0.19\%\ K^{-1}$ (derived from 2006–2010), and observed BCS reported by Brient et al. (2016) of 45% ± 3% (mean ± std, derived from 2006–2012) to obtain observationally-constrained projections accounting for internal variability and observational uncertainty. This calculation was carried out separately for each 20-yr period beginning with 2015-2034 and ending with 2081-2100.

The observational constraint using cloud metrics substantially reduces the uncertainty range for all individual regions relative to the unconstrained uncertainty range from the raw CMIP6 ensembles under both SSP 5-8.5 (Fig 4.6) and SSP 1-2.6 forcing (Fig 4.7). Table AC.S2 provides the 5–95% range and mean projection values for the two scenarios in 2081–2100. Specifically, the application of observational constraints to projections under both SSP 5-8.5 and SSP 1-2.6 results in a decrease in the 95th percentile and an increase in the 5th percentile, with little change in the multi-model ensemble mean projections in all regions. Under rising emissions (SSP 5-8.5, Fig 4.6), the period of most substantial reductions of uncertainty for most regions appears after 2050 as GHG emissions are particularly large at that time in this scenario. This result is consistent with the previously constrained projections of global mean warming Chapter 3). Under strongly declining emissions (SSP 1-2.6, Fig 4.7), our observationally constrained framework still works well to narrow the projected warming range even though global mean temperature stabilizes, again providing more confident regional warming projections.

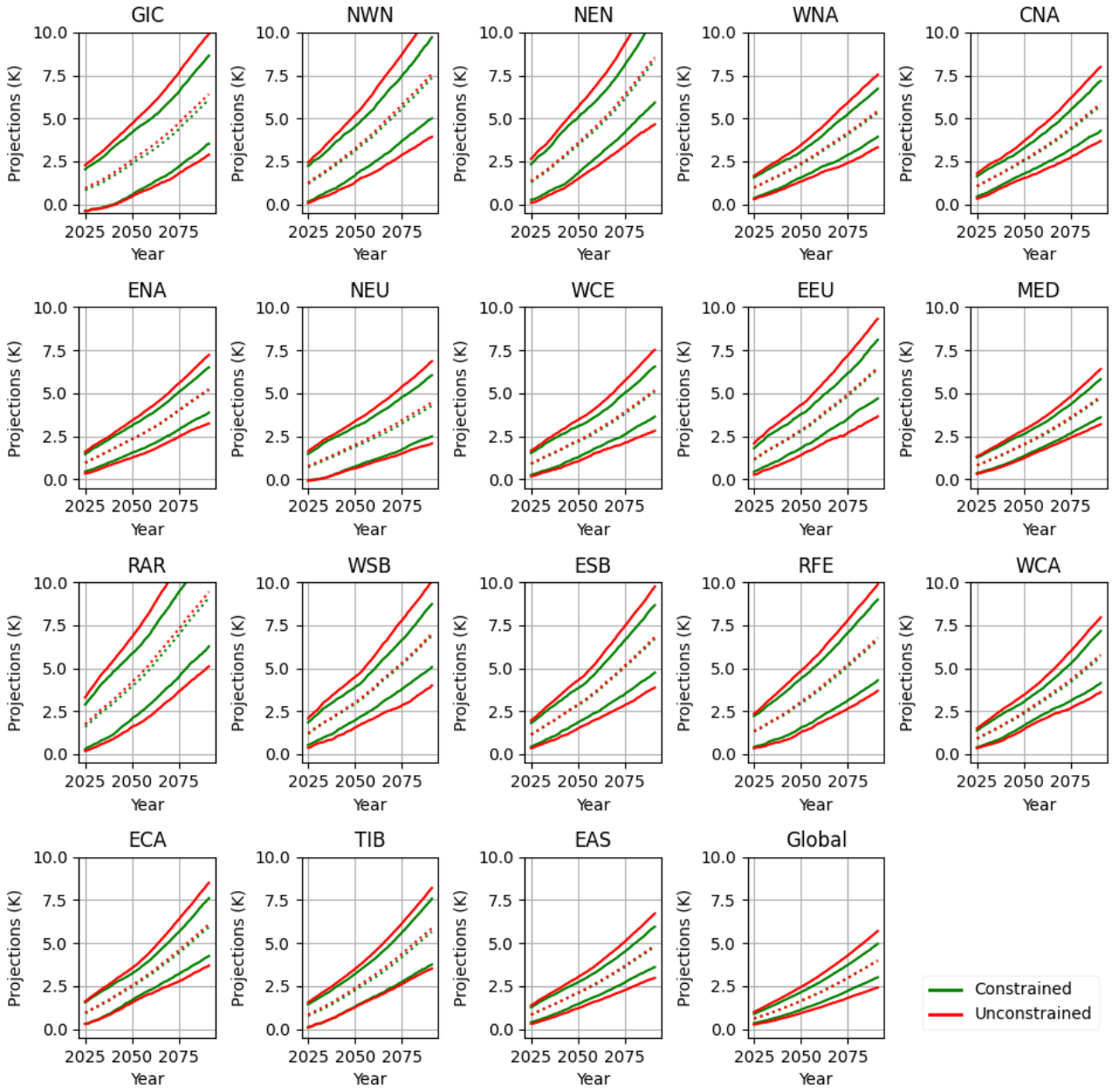


Figure 4. 6 20-year moving average of regional mean near-surface air temperature anomalies (based period:1995-2014) in CMIP6 future projections (SSP5-8.5). The red solid lines show the 5-95% uncertainty range of raw CMIP6 projections and green solid lines show the constrained uncertainty range. The dashed lines show the ensemble mean estimates of constrained and unconstrained projections. The times shown on the *x*-axis are the central years of each 20-year moving average starting from 2015-2034 and ends at 2081-2100.

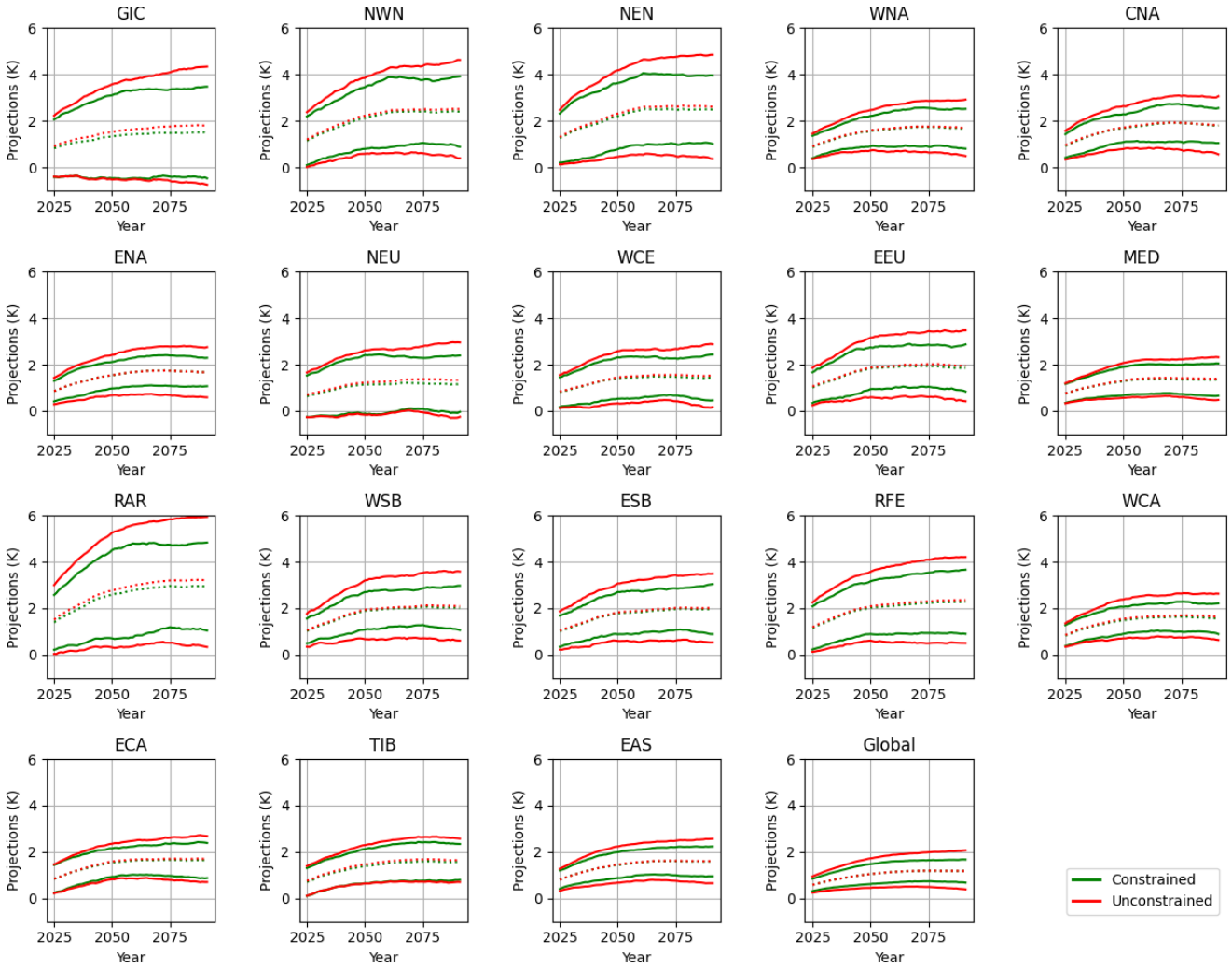


Figure 4. 7 As in Fig 4.6, but for SSP1-2.6.

We present the percentage reductions in late 21st century warming uncertainty for each region and both SSPs in Fig 4.8, corresponding to the observationally constrained results shown in Figs 4.6 and 4.7. The observational constraint reduces the SSP 5-8.5 5-95% range for individual regions by 18% to 40% relative to the unconstrained width, and reduces the SSP 1-2.6 unconstrained 5-95% range by 17% to 44%. For both SSPs, the regional percentage reductions in uncertainty are close to that for the global mean (34%). Given that the relative size of internal variability is expected to be larger on regional scales, this result suggests that uncertainties in the regional constrained projections, like the global mean projections, are dominated by model uncertainty rather than internal variability.

Compared with SSP 5-8.5, most regions involved in this study show somewhat smaller relative reductions (constrained width relative to unconstrained width) in projected temperature change using SSP 1-2.6, with the exceptions of CNA, ENA and RFE. In some regions (GIC, NEU, ECA, TIB and RFE), the effectiveness of the emergent constraint using cloud constraints is worse than the rest of the Northern Hemisphere regions. The relatively weak constraint over these regions may be due to stronger

internal variability or model differences in future regional warming that are more strongly driven by model differences unrelated to tropical and subtropical clouds (e.g. ice albedo feedback and lapse rate feedback, etc.).

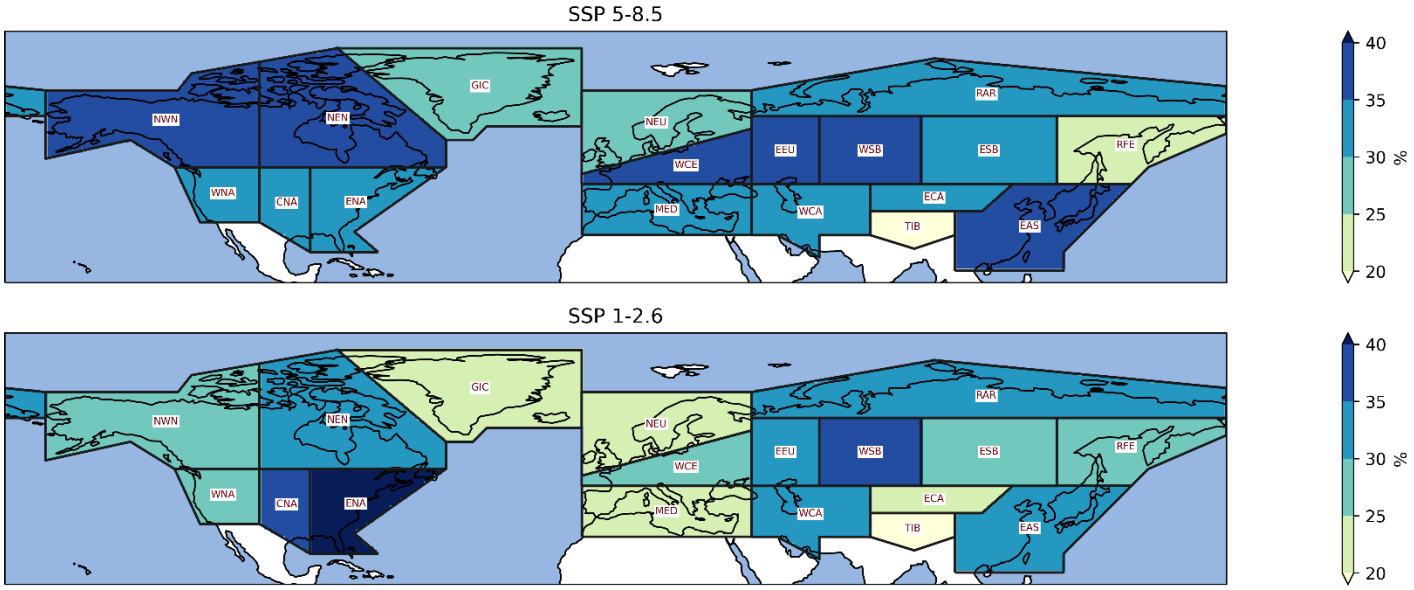


Figure 4. 8 The percentage decrease in the 5-95% uncertainty width due to application of the observational constraints. The constrained and unconstrained widths are derived from Fig. 4.6 and 7 for the 2081-2100 of the 21st century relative to 1995-2014. The upper panel is for SSP 5-8.5 and the lower panel is for SSP 1-2.6.

4.4 Summary and discussion

Other than for Europe (Brunner et al. 2019b; Brunner et al. 2020a; Brunner et al. 2020c), few attempts have been made to observationally constrain projected temperature changes at a regional scale. While IPCC AR6 presented observationally constrained projections of global mean warming, no such constraints were applied to regional warming because 'robust methods do not yet exist to constrain the projections' (Lee et al. 2021).

Using both CMIP5 and CMIP6 simulations, we consider a large number of different regional metrics proposed by previous studies (Brunner et al. 2019a; Lorenz et al. 2018; Senftleben et al. 2020), in addition to a set of global metrics used in previous studies to constrain global temperature changes (Chapter 3; Ribes et al. 2021b; Tokarska et al. 2020). To evaluate and contrast the performance of metrics chosen, we apply a cross-validated imperfect model test. In this evaluation, each model serves in turn as pseudo-observations to constrain projections by all other models, testing how well the future warming in one model can be predicted based on outputs from the other models in the ensemble. The effects of internal variability are accounted for by repeatedly sampling individual members from initial condition ensembles.

Including physically-motivated regional constraints in addition to global ones should improve the constrained projections if the regional metrics provide information to constrain local warming beyond that in the global metrics and are not strongly impacted by internal variability. However, we find that constraining regional temperature projections using regionally-derived metrics introduces the risk of overfitting of the statistical model, and that the resulting projections are more strongly influenced by internal variability. We further show that while the performance of constraints varies across regions, the evaluation process gives us confidence in the accuracy, robustness and reliability of global-scale cloud metrics to observationally constrain regional warming projections. Overall, the cloud metrics result in more robust constrained projections than projections based on regional constraints.

Using observed global-scale cloud metrics we find substantially reduced uncertainties in projected warming for Northern Hemisphere regions compared to unconstrained projections. Our study shows that constraining CMIP6 projections under SSP 5-8.5 and SSP 1-2.6 scenarios results in an increase in the lower bound and a decrease in the upper bound of the warming range with generally little change in mean projected warming. Furthermore, our results show there are reductions of 18-40% for SSP 5-8.5 and 17-44% for SSP 1-2.6 in the uncertainty range of projected warming across regions relative to unconstrained projections.

The accuracy of the constrained projections is generally high over most regions of extratropical Northern Hemisphere, but the accuracy is lower over the TIB region, and the reduction in uncertainty range is smallest there. The relatively poor performance of applying global cloud metrics over this region may be due to model differences in regional warming being driven by local processes that are independent of global warming.

Our study provides regional warming projections which the cross-validated imperfect model test indicates should be accurate, robust, and reliable. These results represent a step forward in metric performance evaluation relative to previous studies applying observational constraints to regional projections. The narrower constrained uncertainty ranges produced by our observationally constrained framework are relevant to climate change policy and adaptation decisions – for example, planning decisions based on the need to avoid particular temperature thresholds.

Chapter 5. Constraining uncertainties in projected warming using the past global warming trend with the pattern effect removed

This chapter will be submitted as:

Liang, Y., Gillett, N. P., & Monahan, A. H. (2023b). Narrowing uncertainties in projected warming by constraining using the past global warming trend with the pattern effect removed (In preparation)

5.1 Introduction and motivation

Future climate change uncertainties based on climate model projections are driven by the choice of different emissions scenarios, uncertainties in the response to external forcing, and internal variability (Deser et al. 2012). An emergent constraint framework based on inter-model correlation between observable metrics in historical simulations and future projected climate provides an effective way to narrow the uncertainty of multi-model projections (Chapter 2&3; Brient and Schneider 2016; Brient et al. 2016; Brunner et al. 2019a; Brunner et al. 2020a; Caldwell et al. 2018; Nijsse et al. 2020; Tokarska et al. 2020; Zhai et al. 2015). Emergent constraints relies on physically-based emergent relations between an observational constraint and the projected warming response to anthropogenic forcing. As an example, past work has highlighted the relevance of simulated warming over recent decades to intermodel variability in simulations of future global warming (Brient and Schneider 2016; Brient et al. 2016; Sanderson et al. 2021). An underlying assumption of an emergent constraint based on the past warming trend is that the emergent relation identified in the models is exchangeable to the real world.

For future long-term global mean warming projections, the past global mean near-surface warming trend can serve as a promising constraint. Since both metrics are based on the response of global mean temperature to increasing GHGs, we would expect a strong correlation across multimodel ensembles given consistent forcing changes from historical to future periods (Nijsse et al. 2020; Sanderson et al. 2021; Tokarska et al. 2020). However, past warming trend constraints are sensitive to the impact of internal variability (Chapter 3; Maher et al. 2020; Po-Chedley et al. 2022; Schwarzwald and Lenssen 2022; Vincent et al. 2015). The past warming trend is therefore not a purely forced response and resulting constrained projections could be biased by the strong influence of internal variability in observations (Chapter3 ; Po-Chedley et al. 2022).

Analysis suggests there is a lower recent warming trend (particularly since 1980s) in observations compared with most CMIP6 model simulations (Tokarska et al. 2020). This fact can be partially understood as the product of internal variability driven by the so-called pattern effect (associated with warming in the western equatorial Pacific

Ocean and cooling in the eastern equatorial Pacific Ocean) (Watanabe et al. 2021; Zhou et al. 2016). Prior studies point out that the relatively cold SST pattern of the eastern tropical pacific (ETP) can exert a substantial negative cloud radiative feedback to reduce global top of atmosphere (TOA) energy budget and GSAT warming (Andrews et al. 2018; Dong et al. 2020; Gregory et al. 2020; Zhou et al. 2016). However, CMIP models generally fail to capture the observed SST pattern over this region (Olonscheck et al. 2020; Fyfe and Gillett 2014). Dong et al. (2022) hypothesize systematic model biases exist over tropical eastern Pacific and the Southern Ocean, and it can be partly due to the lack of realistic meltwater forcing over the Antarctic in the transient climate warming.

However, a single initial condition large ensemble does show a broad range of SST pattern over tropical ocean that can resemble the observed SST pattern (Watanabe et al. 2021). This fact suggests that the relatively cold ETP pattern as observed could be a rare realization of internal variability, which is in principle captured by models with adequate ensemble size. Therefore, the relatively low observed GSAT trend potentially induced by internal variability will favor low climate sensitivity models and tend to result in relatively low constrained warming projections when applying GSAT trend as predictor. This fact motivates some studies to estimate the forced response after removal of internal variability, for use as a constraint on projected warming. For example, Ribes et al. (2022) estimate regional constrained warming over France using the estimated forced response of global and France temperature change as the historical predictor. Po-Chedley et al. (2022) find a relatively large constrained equilibrium climate sensitivity (ECS) using the historical warming trend in the tropical tropospheric region with internal variability removed as a constraint.

Here, we test and apply a framework to separate externally forced warming from unforced internal variability in the past warming trend in both model simulations and observations. First, we identify and regress out the unforced internal variability in the GSAT trend due to the ETP pattern effect. Then, we evaluate and contrast the performance of past GSAT trend as a constraint with and without internal variability reduced. Finally, we contrast and use past GSAT trend with and without internal variability reduced as constraints to estimate projected warming.

5.2 Data and methods

5.2.1 Model data and observation

The predictand in this study is the GSAT change in 2081-2100 relative to 1995-2014 based on CMIP6 model simulations using both SSP 5-8.5 and SSP 1-2.6. This study considers the historical GSAT trend over the period 1970-2022. The historical period 1970-2022 is used for calculating the past warming trend is because the GSAT trend response to aerosol changes over this period is relatively small compared with other longer periods, so that the dominant external forcing is due to greenhouse gases changes

(Chapter 2). In addition, the 53-year period reduces the influence of decadal to multi-decadal internal variability. We also consider the GSAT trend over the relatively shorter period 1993-2012 as a sensitivity analysis to account for a strong influence of internal variability. The period 1993-2012 is selected for calculating the GSAT trend because there is a trend towards cold patterns of SST over the ETP region over this period (Fyfe and Gillett 2014) [ETP is defined as the region east of the dateline and between 20° S and 20° N as in (Kosaka; Xie 2013)]. The CMIP6 model simulations used to calculate GSAT trend (based on near-surface air temperature) and ETP trend (based on SST) are shown in Table AD.S1, while the corresponding observed GSAT trend and ETP trend are derived from HadCRUT5 (Morice et al. 2021) and ERSST V5 (Huang et al. 2017) respectively.

5.2.2 Constrained uncertainty estimates and imperfect model test for evaluation

The emergent constraint on projected warming is obtained using linear regression as described below. We first apply ordinary least-squares regression (OLR) to fit the linear model $\boldsymbol{y} = \alpha + \boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{\beta}$. The historical predictor is denoted by $\boldsymbol{X}$ and future predictand is represented by $\boldsymbol{y}$. The number of rows in $\boldsymbol{X}$ and $\boldsymbol{y}$ is equal to the number M of climate models used. By introducing observational estimates $\boldsymbol{X_0}$ of the predictor into the linear regression model, the best estimate of projected warming due to the constraint is denoted by $\hat{y}_0$. Assuming Gaussian regression error, the future constrained projection is the PDF (Hooper and Zellner 1961; Karpechko et al. 2013; Senftleben et al. 2020)

$$p(y|\boldsymbol{X_0}) = \frac{1}{\sqrt{2\pi\sigma^2_{\hat{y}_0}}} \exp\left(-\frac{(y-\hat{y}_0)^2}{2\,\sigma^2_{\hat{y}_0}}\right) \tag{5.1}$$

where

$$\sigma^2_{\hat{y}_0} = s^2(1 + \boldsymbol{X_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}) \tag{5.2}$$

and

$$s^2 = \frac{1}{M-2}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2 \tag{5.3}$$

We sample one realization from each model in the above analysis to get a PDF based on eqn (5.1). To avoid the under-sampling of internal variability and to account for the contribution of observed uncertainty to the constrained uncertainty, we draw one ensemble member from each model randomly before estimating the regression coefficients, while we sample an observed value drawn from the HadCRUT5 200-member ensemble (Morice et al. 2021) to account for observed uncertainty. Our study does this random process 5000 times to exhaust the choices of internal variability and to account for observed uncertainty. We then average the resulting 5000 distributions to get the final constrained distribution to estimate constrained uncertainty (the justifications of averaging PDFs are shown in Text S1 in Appendix 4).

We apply a cross-validated imperfect model test (Chapter 2) , in which each model serves in turn as pseudo-observations and is used to constrain projections by all other models, to evaluate the performance of constraints. The evaluation is based on values of the root mean square error (RMSE) improvement (relative to the unconstrained ensemble) and correlation coefficient (r), both calculated using the pseudo-observations and the means of the constrained imperfect model ensemble. We also consider probabilistic validation of the imperfect model test (Chapter 2) to assess if the constrained framework can avoid producing overconfident projections. We note whether pseudo-observations lie within the 5-95% constrained uncertainty range for each projection, across all models. Ideally, 90% of the pseudo-observations would lie in the constrained uncertainty in this test.

5.2.3 Removing the unforced internal variability due to ETP SST trend from GSAT trend

There are two steps to remove the internal variability in the GSAT trend that is congruent with variations in ETP SST. Firstly, to obtain a regression relationship between unforced variations in GSAT and unforced variations in ETP, we regress out the difference of forced response over ETP region between models based on the relation between models' ensemble mean ETP trend and models' climate sensitivity (the models' climate sensitivities are represented by ECS, coming from Zelinka et al. (2020)). Secondly, we remove the unforced internal variability in the GSAT trend in observations and each simulation by applying this regression relationship to regress out the unforced ETP variability. The regression model in this step is built up by regressing GSAT trend against ETP trend (with model difference of forced response removed), across all CMIP6 realizations.

5.3 Results

5.3.1 Observed and simulated ETP SST pattern

The SST pattern in the eastern tropical Pacific plays an important role in global mean warming (Andrews et al. 2018; Dong et al. 2020; Gregory et al. 2020; Zhou et al. 2016) . A strong cooling trend was observed over the eastern Tropical Pacific from 1993-2012, as in Fig 5.1 (a). The observed ETP SST trend [(Fig 5.1 (a)] can be reproduced in some individual realizations of CMIP6 models, e.g. the realization 'r7ilp1f2' of MIROC-ES2L. Overall, these models show a broad range of patterns over this region, with either negative or positive ETP SST trends found in different model realizations as the examples shown in Fig 5.1 (b) -(c). Over the 20-years period1993-2012, most model realizations produce a positive warming trend over the ETP, indicating that the observed trend is relatively unlikely from the perspective of the CMIP6 simulation spread (Figure

5.1d). As shown in the uncertainty ranges in Fig 5.1 (d), individual models with large ensembles can largely cover the CMIP6 model spread (with either negative or positive ETP SST trends in a single model), suggesting the internal variability is an important contributor to the variability of warming trend over ETP region. Compared to a similar analysis but based on CMIP5 realizations conducted by Fyfe and Gillett (2014), CMIP6 realizations exhibit a wider range of ETP trends and more CMIP6 realizations are consistent with observed ETP trend. This indicates the larger number of initial-condition ensembles in CMIP6 can provide broader decadal variability to capture observed ETP trend. Also, as assessed in Chapter 3 of IPCC AR6 (Eyring et al. 2021), decadal variability over the Pacific Ocean in CMIP6 models indicate an improved performance (both for spatial structure and magnitude) than CMIP5. Both facts shows that the increased ensemble size of CMIP6 allows for it to better resemble observations, suggesting that the ETP cold pattern in observations is probably a rare realization of internal variability that can be consistent with the modelled ETP trends. Notably, another interpretation for this modelled-observed difference is due to the systematically simulated error. For example, the lack of Southern Ocean cooling due to the forcing of Antarctic meltwater are hypothetically exist in the transient climate simulations, results in relatively warm SST pattern over ETP region (Dong et al. 2022). Our study focuses on the role of internal variability in observed and simulated ETP trend and assess its' influence on GSAT trend (outlined next).

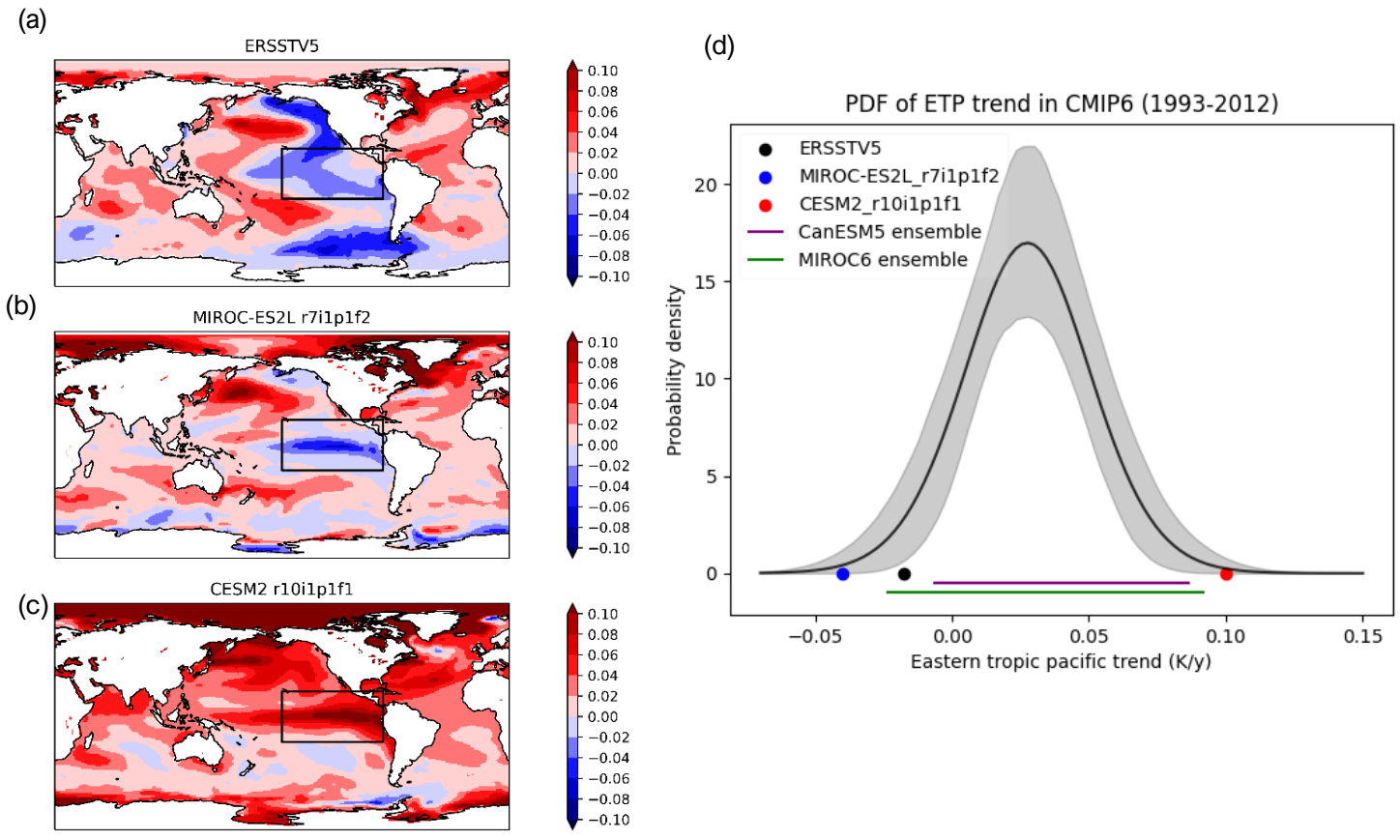


Figure 5. 1 Observed and simulated patterns of SST trends over the period 1993-2012. Panel (a-c) show patterns of SST trend in observation, the ensemble member 'r7ilp1f2' of MIROC-ES2L and the ensemble member 'r10ilp1f1' of CESM2, respectively. Panel (d) shows the PDF of ETP trend in the CMIP6 multi-model ensemble. The PDF is drawn by randomly sampling one random ensemble per model 5000 times (by assuming Gaussian distribution of CMIP6 multi-model ensemble). The solid black curve is the mean of the 5000 samples while the grey shading shows the minimum to maximum range of these 5000 samples. The colored dots in panel (d) correspond to Panel (a), (b) and (c) respectively. The horizonal bars show the ensemble range (minimum to maximum) of trend for individual model CanESM5 and MIROC6.

5.3.2 GSAT trend with the impact of the unforced ETP internal variability removed

We remove the contribution of ETP-related internal variability to the historical GSAT trend in 1970-2022 as well as 1993-2012. The removal procedure includes two major steps, and assumes that the observed and modelled trends are drawn from the same distribution of internal variability.

The first step is to estimate unforced internal variability over ETP region for the set of model realizations. As in Fig AD.S1, the ensemble mean of ETP trend is moderately correlated with ECS across CMIP6 models (correlation coefficient $r = 0.45$, $p = 0.04$). This connection indicates that part of the variance of ETP trend across the multi-model ensemble can be explained by model differences in forced response. Assuming no bias between observed climate sensitivity and the mean ECS of CMIP6 models, we estimate the unforced ETP trend [*x-axis* in Fig 5.2(a)] at the mean value of models' ECS by regressing out the model differences in forced response based on the inter-model relationship between ensemble mean of ETP trend and ECS (Fig AD.S1).

As shown in Fig 5.2(a), the variability of the GSAT trend is tightly connected with the unforced ETP trend (r=0.52, p = 0), indicating that part of GSAT variance can be explained by internal variability from ETP region. The positive correlation reveals that a warm ETP pattern tends to increase the GSAT trend and a cold ETP pattern tends to decrease it, consistent with the physical interpretation of pattern effect discussed in the literature (Dong et al. 2020; Kosaka and Xie 2013; Zhou et al. 2016). Fig 5.2(a) also shows that the observed trend falls within the distribution of simulated trends, though the observed ETP trend is rather small compared with most model realizations.

The second step is to remove unforced internal variability in GSAT trend associated with ETP influence. Based on the regression relationship identified in Fig 5.2(a), we estimate the GSAT trend at the mean value of ETP trend by regressing out the variability of GSAT trend due to unforced ETP internal variability. As shown in Fig 5.2(b), the observed GSAT trend with unforced ETP trend removed approaches the median of the model simulations. The observed GSAT trend increases (red vertical line relative to

blue vertical line) due to the removal of the relatively cold impact arisen from ETP region [Fig 5.2(a)]. In addition, the variance of simulated GSAT trend is slightly reduced after removing unforced internal variability [the red PDF relative to the blue PDF in Fig 5.2(b)].

Repeating these calculations using 1993-2012 as the period for GSAT trend (Fig AD.S2 and Fig AD.S3), we find similar results to those described above.

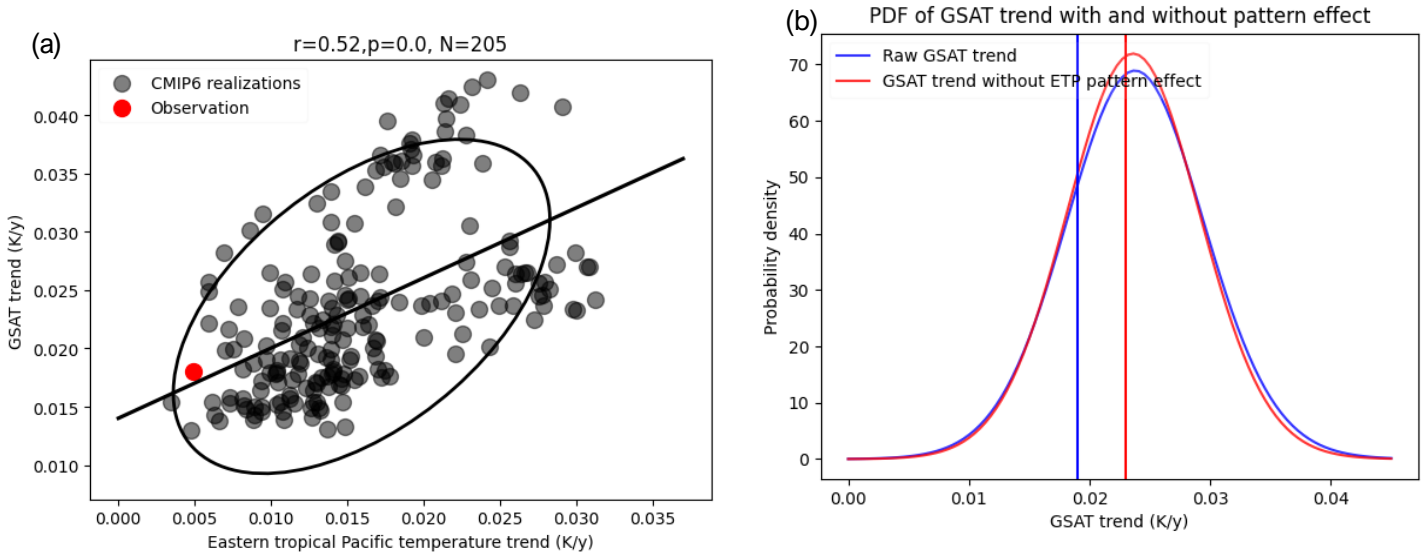


Figure 5. 2 Removing the influence of unforced ETP internal variability on GSAT trend. Panel (a) indicates the relation between the raw GSAT trend and ETP trend (with the forced response removed; mentioned in Fig AD.S1 and Section 5.3.2). The black regression line is estimated from all model realizations (black dots). The black ellipse contains 90% of the probability of the joint distribution (assuming a Gaussian distribution). Panel (b) indicates simulations (shown in PDFs) and observations (shown in vertical line) of raw GSAT trend (in blue lines) and of GSAT trend with ETP SST trend removed (in red lines).

5.3.3 Performance of GSAT trend as a constraint

Removing internal variability in GSAT trend from unforced ETP variability is hypothesized to enhance the emergent relationship between predictor and predictand. The stronger emergent relationship is expected because the signal of forced response is extracted in the historical GSAT trend which should be well correlated with future long-term projected warming. As shown in Fig 5.3, the 1970-2022 GSAT trend (same results but for 1993-2012 are shown in Fig AD.S4) with the pattern effect removed is more correlated with future warming in the global average [Fig 5.3(a)] and in each grid box [Fig5. 3 (b)], compared with the raw GSAT trend as the predictor.

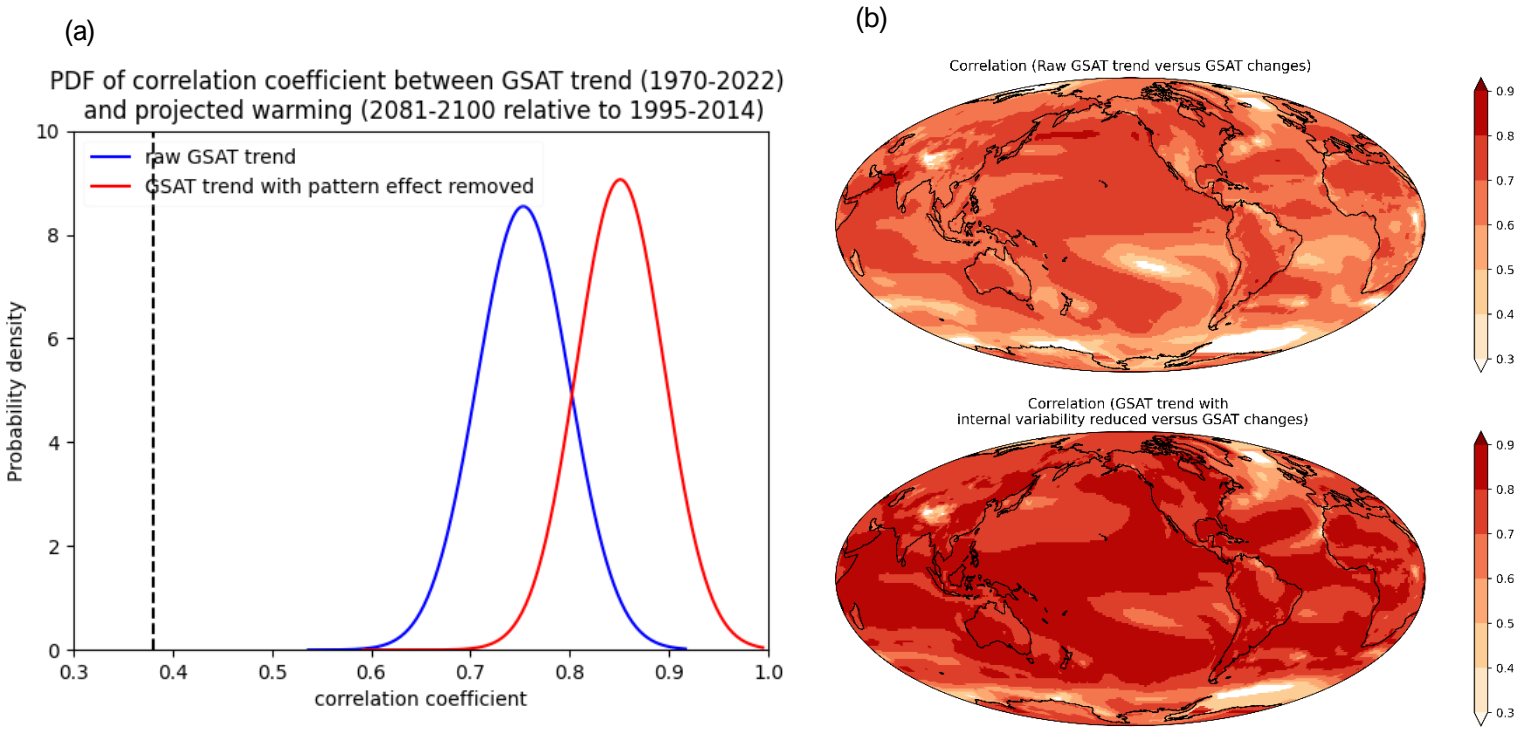


Figure 5. 3 Intermodel correlation between GSAT trend (based on 1970-2022) and projected warming under SSP 5-8.5 (based on 2081-2100 relative to 1995-2014). Panel (a) shows the correlation between historical GSAT trend and projected GSAT changes aross all models. Correlations are based on 5000 random selections of one ensemble per model. Panel (b) shows the intermodel correlation between GSAT trend and projected warming of each grid box (an average across 5000 random selections). The black dashed line in the left panel plot shows the critical value of correlation coefficient that is significant at the 0.05 level. The white shading in panel (b) represents correlation coefficients are not significant at the 0.05 level.

To assess which predictor is more predictive for constrained projections and assess if the predictor chosen produces overconfident projections, we also apply a cross-validated imperfect model test. Our results from the imperfect model test show that constrained projections using the GSAT trend with unforced internal variability reduced result in more accurate mean projections relative to pseudo-observations (red solid lines versus blue solid lines; Fig 5.4 a,b) and a narrower constrained 5-95% uncertainty range than applying raw GSAT trend as a predictor (Fig 5.4 c). A probabilistic validation (Fig 5.4 d) shows no evidence of overconfident constrained projections by applying GSAT trend metrics, because around 90% percent of pseudo-observations lie in the constrained uncertainty range resulting from both predictors.

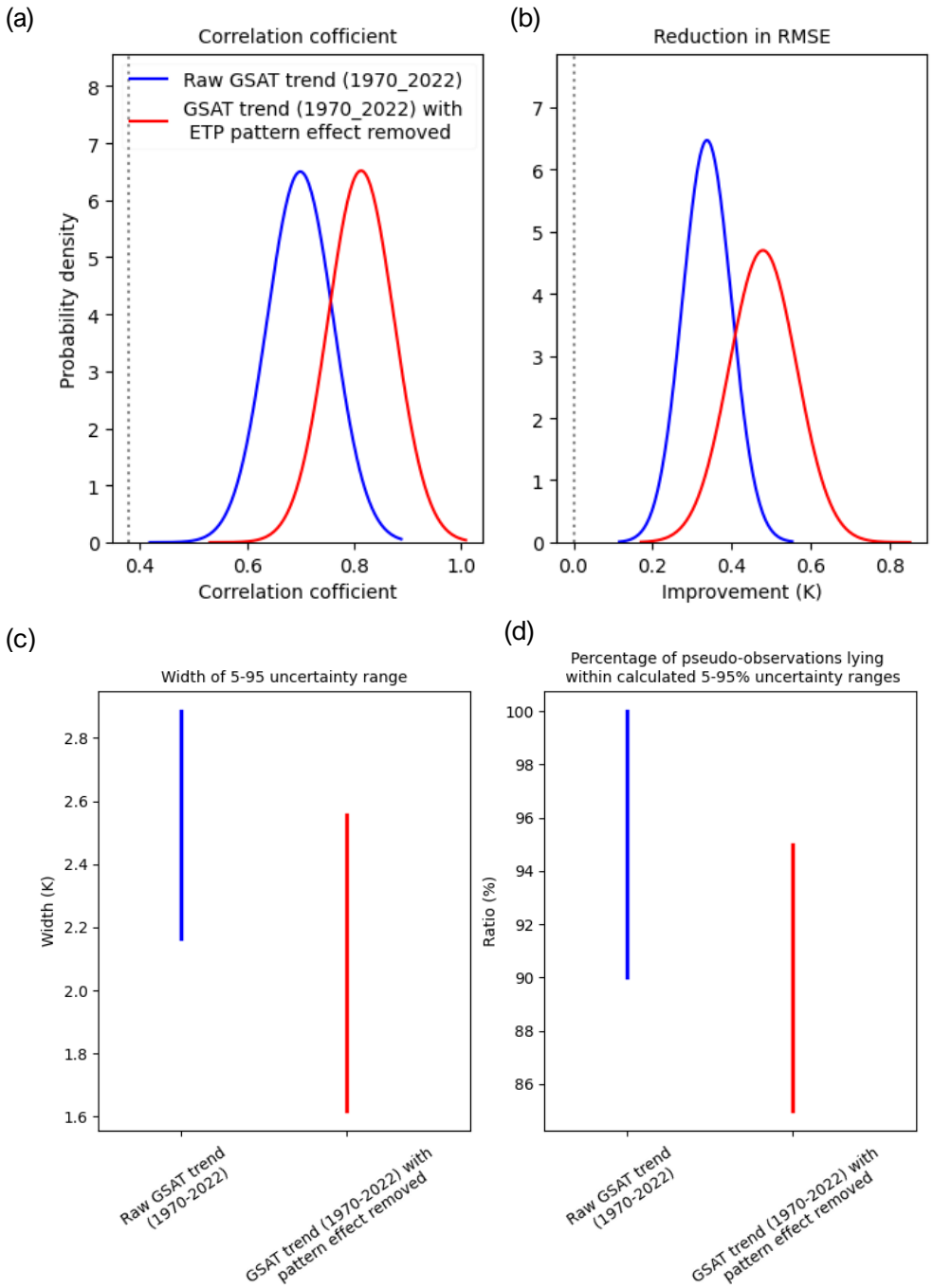


Figure 5. 4 Imperfect model test of constrained projections of warming in 2081-2100 relative to 1995-2014 under SSP5-8.5. Panel (a) shows distributions of correlation coefficient between pseudo-observations and constrained projected mean. Panel (b) shows distributions of RMSE reduction calculated as the RMSE of the unconstrained mean relative to pseudo-observations minus the RMSE of the constrained mean relative to pseudo-observations. The distributions in Panel (a) and (b) are based on 5000 random samples of one ensemble member from each model. The black dashed line in panel (a) shows the critical value of correlation coefficient that is significant at the 0.05 level. The black dashed line in panel (b) shows the threshold 0 representing no improvement from due to the constraint. Panel (c) shows the width (95% percentile minus 5%

percentile) of the constrained uncertainty range. The bars show the 5-95th percentile range of 5000 random samples. Panel (d) shows an evaluation of the reliability of constrained uncertainty. The bars show 5-95th percentiles of 5000 random draws of one ensemble member per model from the ensembles.

5.3.4 Observationally constrained future GSAT changes

We now apply the observed predictor to constrain warming projections using the linear regression approach described in Section 5.2.2. The period 1970-2022 is chosen as the optimal period for calculating predictor GSAT trend due to its strong correlation with projected warming and relatively long duration, which reduces the impact of internal variability (Text S2 in Appendix 4).

For constrained projections based on SSP 5-8.5, the raw GSAT trend metric constrains the 5-95% range to 2.38- 4.73 K while the range constrained by the GSAT trend metric with the pattern effect removed is 3.2-5.07 K, relative to for the unconstrained ensemble 2.34- 5.81K (Fig 5.5). For constrained projections based on SSP 1-2.6, the raw GSAT trend metric constrains the 5-95% range to 0.32- 1.62 K while the range constrained by the GSAT trend metric with pattern effect removed is 0.73-1.87 K, relative to for the unconstrained ensemble 0.37- 2.04 K (Fig AD.S5).

The GSAT trend metric without the unforced ETP variability results in a greater constrained mean, larger constrained 5% and 95% bound, and reduced 5-95% uncertainty range compared to projections constrained using the raw GSAT trend metric. The constrained projections using the GSAT trend without the unforced ETP variability indicate closer agreement with the projections constrained based on cloud metrics (Chapter 3). As for the constrained width of 5-95% range, due to the stronger emergent relation arising from the reduction of unforced internal variability associated with the ETP pattern effect in the historical GSAT trend, a narrower uncertainty range is estimated relative to the constrained uncertainty using the raw GSAT trend metric.

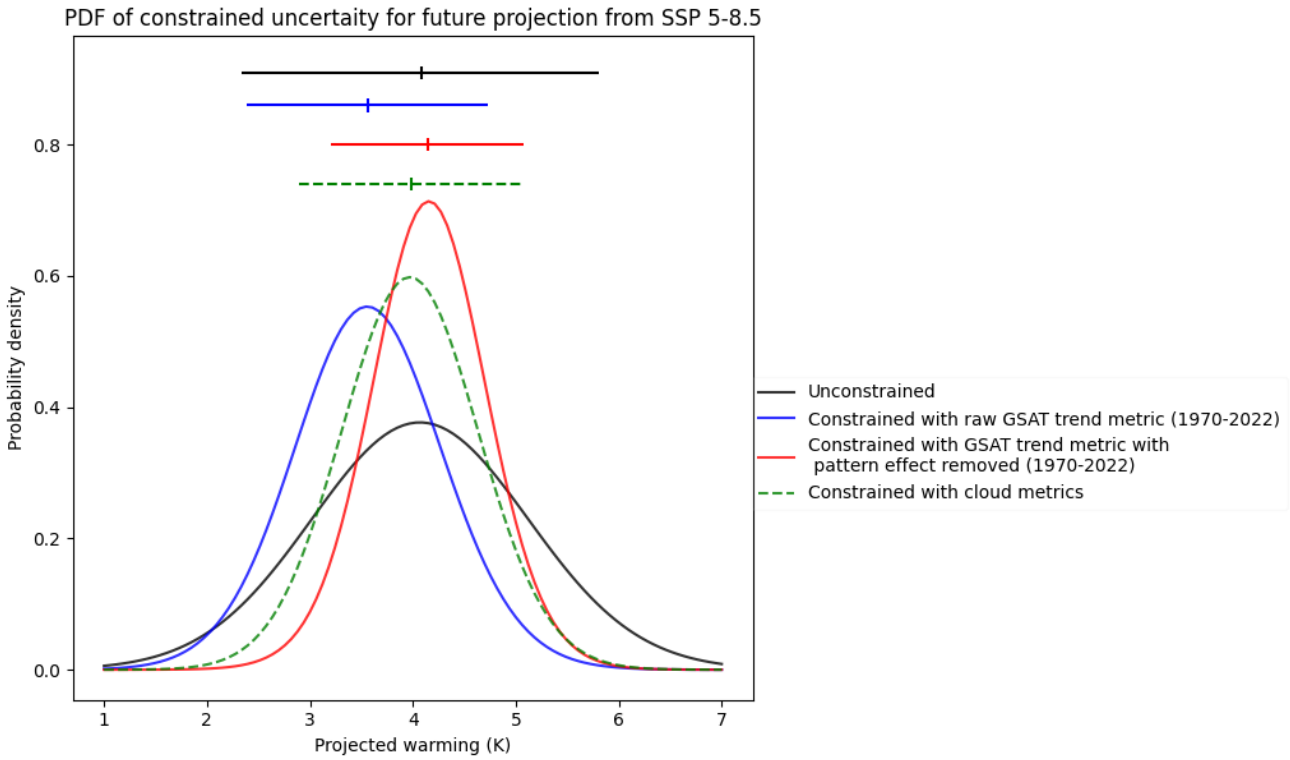


Figure 5. 5 PDFs of constrained and unconstrained projected GSAT changes in 2081-2100 (SSP 5-8.5) relative to 1995-2014. The blue curve shows constrained projections using the raw GSAT trend, while the black curve shows the unconstrained projections. The red curve shows constrained projections using the GSAT trend with the ETP unforced internal variability removed. The green dashed curve shows constrained projections using climatological cloud metrics from Chapter 3. The distributions are generated by sampling over internal variability and observational uncertainty as described in Section 5.2.2.

5.4 Discussion and Conclusion

The warming trend since the 1970s has been dominated by the effects of greenhouse gas increases and simulated warming trends over this period are closely correlated with future warming across models. Such relationship enables us to apply historical GSAT trend as a predictor to constrain projected warming over the 21st century. However, the GSAT trend obtained from the historical period is sensitive to internal variability, weakening the emergent relationship with future warming. A strong cold SST trend in Eastern Tropical Pacific was observed between 1993 and 2012, resulting in a reduction of the GSAT trend due to the so-called pattern effect. Much of the difference between simulated and observed historical GSAT trends are due to the associated unforced internal variability.

We identify a large variability of ETP trend across model simulations and a strong

connection between ETP and GSAT warming in CMIP6 model simulations. Hence, to produce a more robust emergent constraint, we remove the variability in the GSAT trend correlated with unforced variability in the ETP. The observed GSAT trend over the 1970-2022 period with unforced internal variability removed is close to the CMIP6 multi-model mean trend, in contrast with the fact that most raw model simulations overestimate the observed GSAT trend over this period. Assessing the emergent relationship with an imperfect model test, we find improved projection skill by applying past GSAT trend with the unforced internal variability removed.

Applying emergent constraints to projected warming using the observed historical GSAT trend as a constraint, we find a stronger constrained projected warming and a narrower uncertainty range using GSAT trend without unforced internal variability (due to ETP variability) compared with constrained projections using the raw GSAT trend. Compared with unconstrained 5-95% uncertainty range of 2081-2100 warming based on SSP5-8.5, the constrained 5-95% uncertainty range is reduced by 46% based on the metric of 1970-2022 GSAT trend with internal variability reduced, by 38% based on climatology cloud metrics, and by 32% based on the metric of raw 1970-2022 GSAT trend. Using the metric of 1970-2022 GSAT trend with internal variability reduced results in 22% narrower 5-95% uncertainty from IPCC AR6 assessed range based on observational constraint approach [Chapter 4 of AR6. (Lee et al. 2021)]. The constrained projections using the GSAT trend with internal variability reduced is more consistent with the results using climatological cloud metrics as constraint (Chapter 3), indicating that predictors with reduced influence of internal variability result in more robust constrained projections.

Using a relatively longer period (e.g. 1970-2022 relative to 1970-2014 in Chapter 2) for the calculation of historical GSAT trend can reduce the influence of internal variability in emergent relationship, which can result in a narrower constrained uncertainty range. However, even for the relatively long period (e.g. 1970-2022), decadal and multi-decadal internal variability cannot be ignored with respect to forced response at historical period. This fact reflects on the 21% narrower constrained uncertainty range using 1970-2022 GSAT trend (with reduced internal variability) relative to using raw 1970-2022 GSAT trend, for 2081-2100 warming projection based on SSP5-8.5.

It should be noted that inherent advantages of GSAT trend metric exist for use as an emergent constraint. The constraint of GSAT trend is expected to combine all the individual feedbacks for the GSAT warming response, whereas the emergent constraint using metrics relying on the single type of feedback mechanism (e.g. cloud feedback related- cloud metrics) should assume all other feedback are unbiased. This assumption for single line of evidence is likely to produce overconfident constrained projections, which should be evaluated based on a set of additional cross-validated test (as in Chapter 3). This study assumes the exchangeability between models' simulations and the real world. If models have a systematic bias [e.g. due to misrepresented Antarctic

meltwater (Dong et al. 2022)], it could be another caveat for improving the confidence of future constrained projections. In addition, our approach relies on the good representation of models in capturing the relationship between tropical Pacific SST pattern and TOA energy budget and GSAT warming.

Our study provides a framework to reduce the influence of unforced internal variability in emergent constraints of future warming. The past warming trend for the constrained projections will benefit from the longer historical record. Our results imply that the future climate warming inferred by past warming trend in previous studies (e.g. Chapter 2) may be biased low due to the unforced internal variability in observations.

Chapter 6. Summary and Conclusions

Emergent constraints are based on relationships between projected warming and observable features of simulated past climate, across climate models. Such a framework can be used to derive constrained projections of future warming, which have narrower uncertainties than those derived directly from an ensemble of climate models. This framework is highly policy-relevant, and it has been widely used in the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6) to provide robust future warming projections.

The work presented in this dissertation addressed four key aspects of the emergent constraint global and regional warming. The key conclusions, their significance, and contributions to advancement of knowledge resulting from each project are summarized in the Section 6.1. The final Section 6.2 provides overall summary and directions for future research.

6.1 Summary and Significance of Key Findings

6.1.1 Chapter 2: Summary and significance of key findings

The 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.

Chapter 2 first evaluates 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 to projected warming using the observed 1970–2014 GSAT trend as a constraint, Chapter 2 finds 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 large-forcing scenarios SSP3-7.0 and SSP5-8.5, the range of the 95th percentile warming across single-member per model samples is substantially reduced by weighting, relative to the unweighted range.

6.1.2 Chapter 3: Summary and significance of key findings

Chapter 3 indicates that the projected warming is well correlated with tropical and

subtropical low-level cloud properties. These physically meaningful relations enable us to use observed cloud properties to constrain future climate warming. Chapter 3 develops multivariate linear regression models with metrics selected from a set of potential constraints based on a stepwise selection approach. The resulting linear regression model using two low-cloud metrics shows better cross-validated results than regression models that use single metrics as constraints.

Application of a regression model using the low-cloud metrics to constrainclimate projections results in similar estimates of the mean, but substantially narrower uncertainty ranges, of projected twenty-first-century warming when compared with unconstrained simulations. Chapter 3 provides evidence for a higher lower bound of the projected warming range than that obtained from constrained projections based on the past global-mean temperature trend. Consideration of the impact of the sea surface temperature pattern effect on the recent observed warming trend, which is not well captured in the CMIP6 ensemble, indicates that the relatively low projected warming resulting from the global-mean temperature trend constraint may not be reliable and provides further justification for the use of climatologically based cloud metrics to constrain projections.

6.1.3 Chapter 4: Summary and significance of key findings

Few attempts have been made to apply emergent constraint at the regional scale where uncertainties are large. Chapter 4 develops and applies multivariate linear regression models for the projection of surface air temperature averaged over sub-continental regions in the extratropical Northern Hemisphere, based on a set of potential constraints including tropical and subtropical low-level cloud metrics as well as a series of regional climate metrics previously used in the literature.

Based on cross-validated evaluation and compared with unconstrained projections, the projections constrained using low-level cloud metrics exhibit more accurate best estimate projections, narrower uncertainty ranges and more reliable uncertainty estimates in most Northern Hemisphere regions except the Tibetan Plateau. Compared with unconstrained projections, application of the approach to climate projections based on Shared Socioeconomic Pathway (SSP) 1-2.6 and SSP 5-8.5 using observed low-cloud metrics results in considerably narrower 5-95% uncertainty ranges of 21st-century warming over sub-continental Northern Hemisphere land regions.

6.1.4 Chapter 5: Summary and significance of key findings

There is large unforced internal variability of sea surface temperature (SST) trend over the eastern tropical pacific (ETP) which is well correlated with the global warming trend. The strong cooling in the ETP in observations induces a global-scale cooling, yet most realizations in the CMIP6 multi-model ensemble cannot reproduce it.

By removing the unforced internal variability associated with variation in the ETP in observed and simulated GSAT trends, Chapter 5 finds an enhanced correlation between GSAT trends and projected warming and improved results in an imperfect model test. Chapter 5 shows a relatively higher 21st century warming than a constrained projection based on the raw GSAT trend, and brings constrained projections into much closer agreement with projections constrained using climatological cloud metrics.

6.2. Synthesis of Results and Future Directions

6.2.1 Synthesis of Results

The work presented in this dissertation addressed key aspects of constraining model uncertainty of projected warming, focusing on the role of internal variability as well as the performance of historical metrics in constrained projections. Overall, this study demonstrates a path to substantially narrowing uncertainties in projected warming. This study has thoroughly investigated many aspects of constraint, including sensitivity of constrained projections to the constraint method, the choice of constraints, the removal of internal variability in constraint and application on global and regional scales. This study ends up with a narrower range of projected warming at both regional and global scales.

My study builds up an evaluation framework called cross validated test to assess the performance of emergent constraint methods chosen as well as the performance of historical metrics in use. The evaluation framework accounts for internal variability by sampling over different realizations, to assess the important role of internal variability in historical constraint for constraining process. Our study also provides a framework to reduce the influence of unforced internal variability, to provide a strong emergent relationship.

The fundamental contribution that my work has provided is that harmonized projected warming ranges can be obtained even using very different constraints when internal variability is properly accounted for. Previous studies do show a range of constrained projections by applying different constraints. For example, models with strongly positive low-cloud feedback are more consistent with observed constraint related to low-cloud feedback than models with weakly positive or negative feedback, suggesting a relatively higher constrained warming than unconstrained (Brient and Schneider 2016; Vial et al. 2013; Zhai et al. 2015). By contrast, studies focusing on past warming trend as well as Earth's energy budget generally point toward a lower constrained warming (Otto et al. 2013; Nijsse et al. 2020; Tokarska et al. 2020). My study investigates the influence of internal variability in the past warming trend as constraint and our results imply that the future climate warming inferred by past warming trend in previous studies may be biased low due to the unforced internal variability in observations. Consistent constrained projections are found using physically reasonable cloud constraints (with overconfidence issues properly assessed) as well as past warming trend with internal variability reduced.

The evaluation of constrained process and the consistency by applying lines of evidence (e.g. past warming trend with internal variability reduced and climatology cloud metrics) in emergent constraint justify the confidence of very narrow constrained uncertainty range of 21st century warming that this study provides.

In general, these four studies confirm the robustness by accounting for internal variability in predictors and wide applicability of observational constraint framework, and its policy-relevant implications for limiting global and regional mean warming to a given emission scenario. The refined evaluation framework developed in this research can provide some reference for the assessment of climate projections, to better understand the reliability of applying a variety of constrained approaches and potential constraints. The constrained projections produced by this research can feed adaptation planning for a series of activities and provides justification for urgent climate action.

6.2.2 Future directions

Further work will focus on regional warming constraints. Involving regional metrics with regionally physically-motivated basis and the use of GSAT trend with internal variability reduced (e.g. GSAT trend with ETP variability removed), could perhaps produce more accurate and robust constrained projections for regional constraints. Future works will consider the development and improvement to the conceptional framework used to apply emergent constraint (e.g. plan to use total least squares regression as one of constraining method to deal with the issue of internal of variability in predictor). Future works will also benefit from upcoming CMIP7 by models with more initial-condition ensembles, to better estimate the contribution of forced response and internal variability in historical predictors. In addition, as the observed record gets longer and longer in the future, my study expects constrained warming to get narrower and narrower.

Bibliography

Allen, M. R., P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, **407,** 617-620.

Andrews, T., and Coauthors, 2018: Accounting for Changing Temperature Patterns Increases Historical Estimates of Climate Sensitivity. *Geophysical Research Letters*, **45,** 8490-8499.

Bindoff, N. L., and Coauthors, 2014: Detection and Attribution of Climate Change: from Global to Regional. *Climate Change 2013: the Physical Science Basis*, T. F. Stocker, and Coauthors, Eds., Cambridge Univ Press, 867-952.

Bracegirdle, T. J., and D. B. Stephenson, 2012: Higher precision estimates of regional polar warming by ensemble regression of climate model projections. *Climate Dynamics*, **39,** 2805-2821.

Bretherton, C. S., 2015: Insights into low-latitude cloud feedbacks from high-resolution models. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, **373**.

Bretherton, C. S., and P. M. Caldwell, 2020: Combining Emergent Constraints for Climate Sensitivity. *Journal of Climate*, **33,** 7413-7430.

Bridgman, H. A., and J. E. Oliver, 2006: Global Climate System: Patterns, Processes, and Teleconnections. *Global Climate System: Patterns, Processes, and Teleconnections*, 1-331.

Brient, F., and T. Schneider, 2016: Constraints on Climate Sensitivity from Space-Based Measurements of Low-Cloud Reflection. *Journal of Climate*, **29,** 5821-5835.

Brient, F., T. Schneider, Z. H. Tan, S. Bony, X. Qu, and A. Hall, 2016: Shallowness of tropical low clouds as a predictor of climate models' response to warming. *Climate Dynamics*, **47,** 433-449.

Brunner, L., R. Lorenz, M. Zumwald, and R. Knutti, 2019a: Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, **14**.

——, 2019b: Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, **14,** 10.

Brunner, L., A. G. Pendergrass, F. Lehner, A. L. Merrifield, R. Lorenz, and R. Knutti, 2020a: Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, **11,** 995-1012.

——, 2020b: Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth Syst. Dynam. Discuss.*, **2020,** 1-23.

Brunner, L., and Coauthors, 2020c: Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework. *Journal of Climate*, **33,** 8671-8692.

C3S, C. C. C. S., 2017: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. d*ate of access*. https://cds.climate.copernicus.eu/cdsapp#!/home (accessed March 2019). *C. C. C. S. C. D. S. (CDS)*, Ed., *date of access*. https://cds.climate.copernicus.eu/cdsapp#!/home (accessed March 2019).

Caldwell, P. M., M. D. Zelinka, and S. A. Klein, 2018: Evaluating Emergent Constraints on Equilibrium Climate Sensitivity. *Journal of Climate*, **31,** 3921-3942.

Caldwell, P. M., M. D. Zelinka, K. E. Taylor, and K. Marvel, 2016: Quantifying the Sources of Intermodel Spread in Equilibrium Climate Sensitivity. *Journal of Climate*, **29,** 513-524.

Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson, 2014: Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, **41,** 1803-1808.

Collins, M., and Coauthors, 2014: Long-term Climate Change: Projections, Commitments and

Irreversibility. *Climate Change 2013: the Physical Science Basis*, T. F. Stocker, and Coauthors, Eds., Cambridge Univ Press, 1029-1136.

Cox, P. M., C. Huntingford, and M. S. Williamson, 2018: Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, **553,** 319-+.

Daines, J. T., A. H. Monahan, and C. L. Curry, 2016: Model-Based Projections and Uncertainties of Near-Surface Wind Climate in Western Canada. *Journal of Applied Meteorology and Climatology*, **55,** 2229-2245.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012: Communication of the role of natural variability in future North American climate. *Nature Climate Change*, **2,** 775-779.

Dong, Y., K. C. Armour, M. D. Zelinka, C. Proistosescu, D. S. Battisti, C. Zhou, and T. Andrews, 2020: Intermodel Spread in the Pattern Effect and Its Contribution to Climate Sensitivity in CMIP5 and CMIP6 Models. *Journal of Climate*, **33,** 7755-7775.

Dong, Y., Pauling, A. G., Sadai, S., & Armour, K. C. (2022). Antarctic ice-sheet meltwater reduces transient warming and climate sensitivity through the sea-surface temperature pattern effect. Geophysical Research Letters, 49, e2022GL101249. https://doi.org/10.1029/2022GL101.

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

Flynn, C. M., & Mauritsen, T. (2020). On the climate sensitivity and historical warming evolution in recent coupled model ensembles. Atmospheric Chemistry and Physics, 20(13), 7829-7842.

Forster, P., and Coauthors, 2021: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Cambridge University Press. In Press.

Forster, P. M., T. Andrews, P. Good, J. M. Gregory, L. S. Jackson, and M. Zelinka, 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.

Fyfe, J. C., and N. P. Gillett, 2014: Recent observed and simulated warming. *Nature Climate Change*, **4,** 150-151.

Gettelman, A., and Coauthors, 2019: High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2). *Geophysical Research Letters*, **46,** 8329-8337.

Gillett, N. P., 2015: Weighting climate model projections using observational constraints. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, **373,** 8.

Gillett, N. P., and Coauthors, 2021: Constraining human contributions to observed warming since the pre-industrial period. *Nature Climate Change*.

Gregory, J. M., T. Andrews, P. Ceppi, T. Mauritsen, and M. J. Webb, 2020: How accurately can the climate sensitivity to CO2 be estimated from historical climate change? *Climate Dynamics*, **54,** 129-157.

Hall, A., P. Cox, C. Huntingford, and S. Klein, 2019: Progressing emergent constraints on future climate change. *Nature Climate Change*, **9,** 269-278.

Hattab, M. W., C. S. Jackson, and G. Huerta, 2019: Analysis of climate sensitivity via high-dimensional principal component regression. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **5,** 394-414.

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. Bulletin of the American Meteorological Society, 90(8), 1095-1108.

Hirschi, M., and S. I. Seneviratne, 2010: Intra-annual link of spring and autumn precipitation over France. *Climate Dynamics*, **35,** 1207-1218.

Hooper, J. W., and A. Zellner, 1961: The Error of Forecast for Multivariate Regression Models. *Econometrica*, **29,** 544-555.

Hu, X. M., J. R. Ma, J. Ying, M. Cai, and Y. Q. Kong, 2021: Inferring future warming in the Arctic from the observed global warming trend and CMIP6 simulations. *Advances in Climate Change Research*, **12,** 499-507.

Huang, B. Y., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *Journal of Climate*, **30,** 8179-8205.

Iturbide, M., and Coauthors, 2020: An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth System Science Data*, **12,** 2959-2970.

Jimenez-de-la-Cuesta, D., and T. Mauritsen, 2019: Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming. *Nature Geoscience*, **12,** 902-+.

Karpechko, A. Y., D. Maraun, and V. Eyring, 2013: Improving Antarctic Total Ozone Projections by a Process-Oriented Multiple Diagnostic Ensemble Regression. *Journal of the Atmospheric Sciences*, **70,** 3959-3976.

Knutti, R., and J. Sedlacek, 2013: Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, **3,** 369-373.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, **40,** 1194-1199.

Knutti, R., J. Sedlacek, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, 2017: A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, **44,** 1909-1918.

Kosaka, Y., and S. P. Xie, 2013: Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, **501,** 403-+.

Lee, J.-Y., and Coauthors, 2021: Future global climate: scenario-based projections and near-term information. *Climate Change 2021 : The Physical Science Basis. Contribution of Working Group I to the Sixth : Assessment Report of the Intergovernmental Panel on Climate Change : Chapter 4*, V. Masson-Delmotte, and Coauthors, Eds., IPCC, 1-195.

Liang, Y., N. P. Gillett, and A. H. Monahan, 2020: Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend. *Geophysical Research Letters*, **47,** e2019GL086757.

Liang, Y., N. P. Gillett, and A. H. Monahan, 2022: Emergent Constraints on CMIP6 Climate Warming Projections: Contrasting Cloud- and Surface Temperature–Based Constraints. *Journal of Climate*, **35,** 1809-1824.

Liang, Y., N. P. Gillett, and A. H. Monahan, 2023a: Observationally-constrained projections of 21st century regional warming in the extratropical Northern Hemisphere. In review at *Journal of Climate*.

Liang, Y., N. P. Gillett, and A. H. Monahan, 2023b: Narrowing uncertainties in projected warming by constraining using the past global warming trend with the pattern effect removed. To be submitted.

Lorenz, R., N. Herger, J. Sedlacek, V. Eyring, E. M. Fischer, and R. Knutti, 2018: Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America. *Journal of Geophysical Research-Atmospheres*, **123,** 4509-4526.

Lukas, B., R. Lorenz, M. Zumwald, and R. Knutti, 2019: Quantifying uncertainty in European climate

projections using combined performance-independence weighting. Environ. Res. Lett.

MacDougall, A. H., N. C. Swart, and R. Knutti, 2017: The Uncertainty in the Transient Climate Response to Cumulative CO2 Emissions Arising from the Uncertainty in Physical Climate Parameters. *Journal of Climate*, **30,** 813-827.

Maher, N., F. Lehner, and J. Marotzke, 2020: Quantifying the role of internal variability in the temperature we expect to observe in the coming decades. *Environmental Research Letters*, **15**.

Manabe, S., and R. F. Strickler, 1964: Thermal Equilibrium of the Atmosphere with a Convective Adjustment. *Journal of the Atmospheric Sciences*, **21,** 361-385.

Masson, D., and R. Knutti, 2011a: Climate model genealogy. *Geophysical Research Letters*, **38**.

——, 2011b: Spatial-Scale Dependence of Climate Model Performance in the CMIP3 Ensemble. *Journal of Climate*, **24,** 2680-2692.

Mauritsen, T., and E. Roeckner, 2020: Tuning the MPI-ESM1.2 Global Climate Model to Improve the Match With Instrumental Record Warming by Lowering Its Climate Sensitivity. *Journal of Advances in Modeling Earth Systems*, **12,** e2019MS002037.

Meehl, G. A., and C. Tebaldi, 2004: More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, **305,** 994-997.

Meehl, G. A., and Coauthors, 2020: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, **6,** eaba1981.

Merrifield, A. L., L. Brunner, R. Lorenz, I. Medhaug, and R. Knutti, 2020: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics*, **11,** 807-834.

Morice, C. P., and Coauthors, 2021: An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set. *Journal of Geophysical Research: Atmospheres*, **126,** e2019JD032361.

Mueller, B., and S. I. Seneviratne, 2012: Hot days induced by precipitation deficits at the global scale. *Proceedings of the National Academy of Sciences of the United States of America*, **109,** 12398-12403.

Myers, T. A., R. C. Scott, M. D. Zelinka, S. A. Klein, J. R. Norris, and P. M. Caldwell, 2021: Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. *Nature Climate Change*.

Nijsse, F. J. M. M., P. M. Cox, and M. S. Williamson, 2020: An emergent constraint on Transient Climate Response from simulated historical warming in CMIP6 models. *Earth Syst. Dynam. Discuss.*, **2020,** 1-14.

O'Neill, B. C., and Coauthors, 2016: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, **9,** 3461-3482.

Olonscheck, D., M. Rugenstein, and J. Marotzke, 2020: Broad Consistency Between Observed and Simulated Trends in Sea Surface Temperature Patterns. *Geophysical Research Letters*, **47**.

Po-Chedley, S., J. T. Fasullo, N. Siler, Z. M. Labe, E. A. Barnes, C. J. W. Bonfils, and B. D. Santer, 2022: Internal variability and forcing influence model–satellite differences in the rate of tropical tropospheric warming. *Proceedings of the National Academy of Sciences*, **119,** e2209431119.

Qu, X., A. Hall, S. A. Klein, and P. M. Caldwell, 2014: On the spread of changes in marine low cloud cover in climate model simulations of the 21st century. *Climate Dynamics*, **42,** 2603-2626.

Ramanathan, V., E. J. Pitcher, R. C. Malone, and M. L. Blackmon, 1983: THE RESPONSE OF A SPECTRAL GENERAL-CIRCULATION MODEL TO REFINEMENTS IN RADIATIVE PROCESSES. *Journal of the Atmospheric Sciences*, **40,** 605-630.

Renoult, M., and Coauthors, 2020: A Bayesian framework for emergent constraints: case studies of

climate sensitivity with PMIP. *Climate of the Past*, **16,** 1715-1735.

Riahi, K., and Coauthors, 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.

Ribes, A., S. Qasmi, and N. P. Gillett, 2021a: Making climate projections conditional on historical observations. *Science Advances*, **7,** eabc0671.

——, 2021b: Making climate projections conditional on historical observations. *Science Advances*, **7**.

Ribes, A., J. Boé, S. Qasmi, B. Dubuisson, H. Douville, and L. Terray, 2022: An updated assessment of past and future warming over France based on a regional observational constraint. *Earth Syst. Dynam. Discuss.*, **2022,** 1-29.

Rieck, M., L. Nuijens, and B. Stevens, 2012: Marine Boundary Layer Cloud Feedbacks in a Constant Relative Humidity Atmosphere. *Journal of the Atmospheric Sciences*, **69,** 2538-2550.

Rougier, J., M. Goldstein, and L. House, 2013: Second-Order Exchangeability Analysis for Multimodel Ensembles. *Journal of the American Statistical Association*, **108,** 852-863.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015a: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *Journal of Climate*, **28,** 5171-5194.

——, 2015b: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. *Journal of Climate*, **28,** 5150-5170.

Sanderson, B. M., M. Wehner, and R. Knutti, 2017: Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, **10,** 2379-2395.

Sanderson, B. M., A. Pendergrass, C. D. Koven, F. Brient, B. B. B. Booth, R. A. Fisher, and R. Knutti, 2021: On structural errors in emergent constraints. *Earth Syst. Dynam. Discuss.*, **2021,** 1-30.

Schlund, M., A. Lauer, P. Gentine, S. C. Sherwood, and V. Eyring, 2020: Emergent constraints on Equilibrium Climate Sensitivity in CMIP5: do they hold for CMIP6? *Earth Syst. Dynam. Discuss.*, **2020,** 1-40.

Schwarzwald, K., and N. Lenssen, 2022: The importance of internal climate variability in climate impact projections. *Proceedings of the National Academy of Sciences*, **119,** e2208095119.

Sellar, A. A., and Coauthors, 2019: UKESM1: Description and evaluation of the UK Earth System Model. *Journal of Advances in Modeling Earth Systems*, **n/a**.

Seneviratne, S. I., and Coauthors, 2010: Investigating soil moisture-climate interactions in a changing climate: A review. *Earth-Science Reviews*, **99,** 125-161.

Senftleben, D., A. Lauer, and A. Karpechko, 2020: Constraining Uncertainties in CMIP5 Projections of September Arctic Sea Ice Extent with Observations. *Journal of Climate*, **33,** 1487-1503.

Sherwood, S., and Coauthors: An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, **n/a,** e2019RG000678.

Sherwood, S. C., S. Bony, and J. L. Dufresne, 2014: Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, **505,** 37-+.

Sherwood, S. C., and Coauthors, 2020: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics*, **58**.

Simpson, I. R., K. A. McKinnon, F. V. Davenport, M. Tingley, F. Lehner, A. Al Fahad, and D. Chen, 2021: Emergent Constraints on the Large-Scale Atmospheric Circulation and Regional Hydroclimate: Do They Still Work in CMIP6 and How Much Can They Actually Constrain the Future? *Journal of Climate*, **34,** 6355-6377.

Stephens, G. L., 1984: THE PARAMETERIZATION OF RADIATION FOR NUMERICAL WEATHER

PREDICTION AND CLIMATE MODELS. *Monthly Weather Review*, **112,** 826-867.

Storch, V., and F. W. Zwiers, 1999: Statistical analysis in climate research. *EBSCO Academic eBook Collection Complete*, Cambridge University Press, x, 484 p.

Stott, P. A., and J. A. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, **416,** 723-726.

Stott, P. A., J. A. Kettleborough, and M. R. Allen, 2006: Uncertainty in continental-scale temperature predictions. *Geophysical Research Letters*, **33**.

Swart, N. C., and Coauthors, 2019: The Canadian Earth System Model version 5 (CanESM5.0.3). Geoscientific Model Development Discussions, 1--68.

Tebaldi, C., and J. M. Arblaster, 2014: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, **122,** 459-471.

Thackeray, C. W., and A. Hall, 2019: An emergent constraint on future Arctic sea-ice albedo feedback. *Nature Climate Change*, **9,** 972-+.

Tokarska, K. B., M. B. Stolpe, S. Sippel, E. M. Fischer, C. J. Smith, F. Lehner, and R. Knutti, 2020: Past warming trend constrains future warming in CMIP6 models. *Science Advances*, **6**.

Vial, J., J. L. Dufresne, and S. Bony, 2013: On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Climate Dynamics*, **41,** 3339-3362.

Vial, J., S. Bony, J. L. Dufresne, and R. Roehrig, 2016: Coupling between lower-tropospheric convective mixing and low-level clouds: Physical mechanisms and dependence on convection scheme. *Journal of Advances in Modeling Earth Systems*, **8,** 1892-1911.

Vincent, L. A., and Coauthors, 2015: Observed Trends in Canada's Climate and Influence of Low-Frequency Variability Modes. *Journal of Climate*, **28,** 4545-4560.

Voldoire, A., and Coauthors, 2019: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, **11,** 2177-2213.

Watanabe, M., J. L. Dufresne, Y. Kosaka, T. Mauritsen, and H. Tatebe, 2021: Enhanced warming constrained by past trends in equatorial Pacific sea surface temperature gradient. *Nature Climate Change*, **11,** 33-+.

Williamson, D. B., and P. G. Sansom, 2019: How Are Emergent Constraints Quantifying Uncertainty and What Do They Leave Behind? *Bulletin of the American Meteorological Society*, **100,** 2571-2588.

Zelinka, M. D., and Coauthors, 2020a: Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, **47,** 12.

——, 2020b: Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, **47**.

Zhai, C. X., J. H. Jiang, and H. Su, 2015: Long-term cloud change imprinted in seasonal cloud variation: More evidence of high climate sensitivity. *Geophysical Research Letters*, **42,** 8729-8737.

Zhang, B., M. Linz, and G. Chen, 2022: Interpreting Observed Temperature Probability Distributions Using a Relationship between Temperature and Temperature Advection. *Journal of Climate*, **35,** 705-724.

Zhou, C., M. D. Zelinka, and S. A. Klein, 2016: Impact of decadal cloud variations on the Earth's energy budget. *Nature Geoscience*, **9,** 871-+.

Zhou, C., M. D. Zelinka, A. E. Dessler, and M. H. Wang, 2021: Greater committed warming after accounting for the pattern effect. *Nature Climate Change*, 6.

Appendix A

This supporting information lists the CMIP6 models (Table AA.S1) and weighting model parameters (Table AA.S2) that we used in this study, the correlation between historical trends and projected warming for different initial years (Table AA.S3), results from imperfect model tests (Table AA.S4, S5), projected values based on the two weighting methods (Table AA.S6), and presents extra text and figures supporting our results.

Table AA.S1 CMIP6 model runs for each Shared Socioeconomic Pathway (SSP) used in this study. For future projections, a complete set of simulations is not available for all models.

		Number of model runs for each scenario				
	Model name	**Historical**	**SSP1-2.6**	**SSP2-4.5**	**SSP3-7.0**	**SSP5-8.5**
1	ACCESS-CM2	1	1	1	1	1
2	ACCESS-ESM1	1	1	1	1	1
3	BCC-CSM2-MR	3	1	1	1	1
4	CAMS-CSM1	2	2	2	2	2
5	CanESM5	50	50	50	50	50
6	CESM2	6	1	1	2	2
7	CESM2-WACCM	3	1	1	1	1
8	CNRM-CM6-1	10	6	6	6	6
9	CNRM-ESM2-1	5	5	5	5	5
10	CNRM-CM6-1-HR	1	1	1	1	1
11	EC-Earth3	3	7	6	7	3
12	EC-Earth3-Veg	3	3	3	3	3
13	FGOALS-f3-L	3	1	1	1	1
14	FGOALS-g3	1	1	1	1	1
15	FIO-ESM-2-0	3	3	3	3	3
16	GFDL-ESM4	1	1	1	1	1
17	HadGEM3-GC31-LL	4	1	1	1	-
18	INM-CM5-0	8	1	1	1	1
19	INM-CM4-8	1	1	1	1	1
20	IPSL-CM6A-LR	16	3	5	10	5
21	KACE-1-0-G	3	2	2	2	2
22	MCM-UA-1-0	2	1	1	1	1
23	MIROC6	9	3	3	3	3
24	MIROC-ES2L	3	1	1	1	1
25	MPI-ESM1-2-HR	5	1	1	5	1
26	MPI-ESM1-2-LR	8	8	8	8	8
27	MRI-ESM2-0	4	1	1	5	1
28	NESM3	5	2	2	-	2

29	NorESM2-LM	3	1	1	1	1
30	UKESM1-0-LL	8	5	5	5	4

Table AA.S2 Weighting model parameters ($\boldsymbol{\sigma_d}$ and $\boldsymbol{\sigma_s}$) for GSAT trend (m1), root-mean-square-difference (RMSD) of gridded SAT (m2) and the compound metric (m3) for different periods used in this study, derived based on ensemble means.

Scenarios	σ_d(m1) of 2081-2100/2041-2060/2021-2040	σ_d(m2) of 2081-2100/2041-2060/2021-2040	σ_d(m3) of 2081-2100/2041-2060/2021-2040
SSP1-2.6	0.52/0.47/0.51	0.38/0.36/0.37	0.46/0.43/0.45
SSP2-4.5	0.58/0.57/0.53	0.40/0.38/0.37	0.45/0.43/0.41
SSP3-7.0	0.61/0.51/0.57	0.39/0.38/0.41	0.42/0.40/0.42
SSP5-8.5	0.72*/0.62*/0.60	0.42/0.43/0.42	0.58*/0.56/0.58

Since σ_s values are calculated for each metric from the historical period, we have a single σ_s for each metric: σ_s (m1)=1.17, σ_s (m2)=0.77, σ_s (m3)=0.81.

* While $\boldsymbol{\sigma_d}$ is generally taken as the lowest value which gives the expected 90% coverage rate in an imperfect model test, in these cases the 90% coverage rate is not met for any value of $\boldsymbol{\sigma_d}$, so we select the minimum σ_d that can get the maximum coverage (the maximum coverages are more than 85% for these cases).

Table AA.S3 Correlation (r) between historical trends and projected warming (2081-2100 versus 1995-2014) under SSP5-8.5 for different trend periods : 1850-2014, 1960-2014, 1970-2014, 1980-2014 and 1990-2014. All p values for correlations listed in the table are less than 0.0001.

Historical period	**1850-2014**	**1950-2014**	**1960-2014**	**1970-2014**	**1980-2014**	**1990-2014**
r	0.43	0.69	0.76	0.80	0.78	0.79

Table AA.S4 Correlation (r) and root-mean-square-error (RMSE) from the imperfect model test computed, based on different metrics, for SSP5-8.5 in 2041-2060 and 2081-2100. Values outside brackets are computed using ensemble means from each model, while values in brackets are means across 5000 single-member per model random samples.

Metric	**r for weighted projections vs pseudo observations for 2041-2060/2081-2100**	**RMSE (units: K) for weighted projections vs pseudo observations for 2041-2060/ 2081-2100**
GSAT trend	0.60** (0.42**)/0.52** (0.44**)	0. 47 (0.37)/1.03 (0.85)
RMSD of gridded SAT	0.54** (0.32*)/0.54** (0.45**)	0.46 (0.38)/0.95 (0.83)
Compound metric	0.57** (0.46**)/0.61** (0.43**)	0.45 (0.37)/0.92 (0.85)

Unweighted	-0.99** (-1.0**)/-1.0 ** (-1.0**)	0.65 (0.42)/1.42 (0.95)

**represent p value is less than 0.05, *represent p value is less than 0.1 but greater than 0.05.

Table AA.S5 Correlation (r) between the mean weighted projection and truth based on imperfect model test for two different metrics for the period of 2041-2060 from SSP5-8.5. We randomly select one member per model to do the imperfect model test. The third column shows the results when we remove one duplicated model for models from a single institution.

Metric	r for pseudo observation vs weighted projection (for all models available)	r for pseudo observation vs weighted projection (removing one duplicated model for models from same institution)
RMSD of gridded SAT	0.32*	0.21
GSAT trend	0.42**	0.41**

**indicates p value is less than 0.05, *indicates p value is less than 0.1 but greater than 0.05

Table AA.S6 Projected mean warming and 5-95% confidence ranges based on the weighting method and unweighted simulations for four SSP scenarios (units: K). The results for random selection on weighted and unweighted are best estimates from the 5000 samples.

Time period	**Scenarios**	**Mean and 5–95th confidence ranges of prediction**				
		Unweighted (one randomly selected member per model)	**Weighted** (with one randomly selected member per model; weighting using GSAT Trend)	**Unweighted** (equal weight for each model)	**Weighted** (weighting using ensmeble mean of each model and GSAT trend)	**Weighted** (weighting using ensemble mean per model and compound metric)
2021-2040	SSP1-2.6	0.71 (0.38, 1.08)	0.68 (0.39, 0.95)	0.71 (0.39, 1.11)	0.67 (0.38, 0.86)	0.67 (0.39, 0.87)
	SSP2-4.5	0.74 (0.40, 1.17)	0.71 (0.41, 0.96)	0.74 (0.40, 1.18)	0.69 (0.39, 0.88)	0.69 (0.38, 0.87)
	SSP3-7.0	0.72 (0.39, 1.14)	0.69 (0.40, 0.97)	0.72 (0.39, 1.18)	0.68 (0.40, 0.94)	0.72 (0.41, 0.94)
	SSP5-8.5	0.83 (0.43, 1.25)	0.80 (0.48, 1.06)	0.83 (0.53, 1.25)	0.79 (0.46, 1.05)	0.80 (0.46, 1.05)

2041-2060	SSP1-2.6	1.02 (0.44, 1.61)	0.96 (0.51, 1.52)	1.02 (0.56, 1.61)	0.92 (0.60, 1.23)	0.93 (0.60, 1.28)
	SSP2-4.5	1.27 (0.80, 1.91)	1.22 (0.83, 1.68)	1.27 (0.82, 1.96)	1.16 (0.83, 1.42)	1.17 (0.83, 1.46)
	SSP3-7.0	1.41 (0.91, 2.18)	1.36 (0.93, 1.83)	1.41 (0.93, 2.24)	1.31 (0.91, 1.81)	1.40 (1.04, 1.80)
	SSP5-8.5	1.69 (1.03, 2.45)	1.64 (1.10, 2.08)	1.69 (1.16, 2.50)	1.60 (1.16, 2.05)	1.62 (1.16, 2.05)
2081-2100	SSP1-2.6	1.11 (0.47, 1.87)	1.03 (0.52, 1.66)	1.11 (0.57, 1.87)	0.96 (0.57, 1.51)	0.97 (0.57, 1.52)
	SSP2-4.5	2.04 (1.31, 2.93)	1.97 (1.32, 2.65)	2.04 (1.31, 2.96)	1.87 (1.37, 2.61)	1.89 (1.37, 2.63)
	SSP3-7.0	3.08 (2.06, 4.57)	3.01 (2.15, 3.91)	3.08 (2.05, 4.69)	2.87 (2.23, 3.81)	3.08 (2.29, 3.75)
	SSP5-8.5	3.90 (2.48, 5.34)	3.82 (2.72, 4.77)	3.90 (2.60, 5.68)	3.68 (2.81, 4.78)	3.75 (2.91, 4.78)

Text S1 Calculation of independence weighting parameter

We apply the method proposed by Brunner (2019a) to estimate the weighting model parameter σ_s for the independence weights ($wd_i = 1 + \sum_{j \neq i}^{M} e^{-\frac{s_{ij}^2}{\sigma_s^2}}$; denominator from Eqn2.1) using the information from models with ensemble size exceeding one. Initially, we use only a single ensemble member of a model to calculate wd_i for a range of σ_s values. We then add the other available members in turn for model j with more than one ensemble member and keep other models still with one member and calculate the independence weights ($\widetilde{wd}_i^{\,ind}$) again. Since ideally all models are independent with each other and all ensemble members of a given model are similar, wd_j will ideally increase by the number N_j of new ensemble members of model j added (measured by $\mu_1 = mean_j\left[wd_j^{ind}(\sigma_s) + N_j - \widetilde{wd}_j^{\,ind}(\sigma_s)\right]^2$), but other models' weights should ideally remain the same (measured by $\mu_2 = m_{ean_j}\left\{mean_i\left[wd_{i\neq j}^{ind}(\sigma_s) - \widetilde{wd}_{i\neq j}^{ind}(\sigma_s)\right]^2\right\}$). We select σ_s to get the minimum error term $\mu = \mu_1 + \mu_2$.

Text S2 Calculation details of the weighting method using the observational constraint

a. Constraining projections with observations using one randomly-selected ensemble member per model

Step1: Randomly select one ensemble member per model. Add the historical trend to the list H1, and future temperature change from the same ensemble member to the list S1 (eg. r1i1p1f1 in both historical and future periods).

Step2: Apply H1 and S1 to determine σ_D from Eqn (2.1) by the imperfect model test. For our targeted σ_D in the range (0-3), 90% of the pseudo-observations should lie in the 5-95% predicted range, and we select the minimum σ_D value that meets this requirement. If 90% coverage cannot be obtained, we use the smallest σ_D with the largest coverage value (Brunner et al. 2019a; Knutti et al. 2017; Lorenz et al. 2018).

Step3: Estimate the Cumulative Distribution Function (CDF) for S1 with weights calculated by Eqn (2.1) with H1 and the observed GSAT trend, and then compute 5%, 95% and mean values from the CDF.

Step4: Repeat Step 1-3 5000 times to build the samples of 5%, 95%, mean value and CDF curve .

b. Constraining projections with observations using the ensemble mean of each model

To minimize the effect of internal variability on model weights and apply all ensemble members for weighting, we also calculate weights using Eqn. (2.1) with the distance measure based on the ensemble mean for each model and then give equal weight to each ensemble member, weighting individual ensemble members by the inverse of the ensemble size. In addition to GSAT trend differences, we also cosnsider weighting based on the RMSD of gridded SAT and a metric which combines temperature trend and RMSD of gridded SAT with equal weight in the distance metric. The method we used to calculate σ_S is described in Text S1. To determine σ_D, we use an imperfect model test similar to that described in Section 2.3, in which we calculate the weights for a range of different σ_D values using each model as a pseudo-observation in turn and examine if the pseudo-observations lie within the 5^{th}–95^{th} percentile range for our target prediction (e.g., GSAT warming in 2081-2100 in SSP5-8.5). We select the minimum value of σ_D for which the coverage ratio is 90% (Brunner et al. 2019a; Knutti et al. 2017; Lorenz et al. 2018). Since σ_D varies between future periods for each SSP, we use different σ_D for different periods and SSPs when computing weights. Table S2 shows the σ_D and σ_S values which we calculate using ensemble means in this study.

Text S3 Calculation on RMSD of gridded SAT

RMSD of gridded SAT metric is calculated by inter-model distance or model-observation distance using gridded near-surface air temperature (SAT). We first interpolate each model to a common resolution of 2.5°×2.5°. Then, the global gridded

SAT from model i, time period 1979-2014, is selected. The climatological mean of the global gridded SAT is calculated over the period 1979-2014. The point-to-point distance from model i to the other model j or the point-to-point distance from model i to the observations is computed and a distance metric d is calculated as the area weighted root-mean-squared difference. Finally, the distance metric based on RMSD of gridded SAT in its final form is normalized by the median over all models.

Text S4 Selection of time period

In order to maximise the strength of the statistical relationship between past trends and future changes in the CMIP6 models, we vary the date range over which trends are calculated. Holding the final year fixed at 2014, the last year in the CMIP6 historical simulation, we calculate the correlation (r) between historical trends and projected warming between 1995-2014 and 2081-2100 under SSP5-8.5 for different initial years : 1850-2014, 1960-2014, 1970-2014, 1980-2014 and 1990-2014. In this calculation we consider all ensemble members from all models. The period 1970-2014 shows the strongest correlation (Table S3) of past trends and future changes.

Also, we randomly pick one member per model and then calculate the correlation based on the simulated GSAT trend and future warming (using the SSP5-8.5 scenario) from the corresponding sample of model members, repeating this operation 5000 times (the corner plot of Figure 2.1a shows the distribution of correlation values.).

Figures S1 (a) and (b) respectively show the inter-model distance matrix and model-observation distance values based on RMSD of gridded SAT. Models from the same institutions show similar gridded SAT climatologies, with relatively small distances (CESM2-WACCM/CESM2, CNRM-ESM2/CNRM-CM6-1, EC-Earth/EC-Earth3-Veg, INN-CM4-8/INN-CM5-0, MPI-ESM1-2-HR/MPI-ESM1-2-LR, and MIROC-ES2L/MIROC6). Based on the comparisons on Figure 1b and Figure S1b, we find no strong relationship between models with good performance at simulating the GSAT trend and those with small RMSD of gridded SAT.

Text S5 Imperfect model test using one randomly-selected ensemble member per model

a. For 2041-2060 and 2081-2100 with SSP5-8.5:

Step1 and Step2: same as Step1 and Step2 in Text S2
Step3: Apply imperfect model test (withholding one model in turn by treating this model as pseudo-observation, and use the remaining models to predict the pseudo-observation) on H1 and S1 by Eqn (2.1), from which correlation coefficient and RMSE between pseudo observations and predicted means can be computed.
Step4: Repeat step 1-3 5000 times to build the samples of correlation coefficient and RMSE.

b. For historical simulation:

Step1: Randomly select one member per model with historical trend to form a list H1.

Step2: Apply H1 to determine σ_D by an imperfect model test (for our targeted σ_D in a range of different σ_D values, 90% pseudo observation should lie in 5-95% predicted range of models, and we select the minimum σ_D value that meet this requirement. If 90% coverage cannot be obtained, we will apply smallest σ_D with the most coverage value).

Step3: Apply the imperfect model test on H1 by Eqn (2.1) to get the correlation coefficient and RMSE.

Step4: Repeat step 1-3 5000 times

Text S6 Imperfect model test using using ensemble means

For the imperfect model test of weighting scheme, each ensemble member will act in turn as a pseudo-observation, and the weighting method is applied using the ensemble mean of each model. Ensemble members from the model whose historical simulation is being used as pseudo-observations are excluded in the calculation.

As expected, when the weighting approach is used to predict historical warming in the simulation treated as pseudo-observations, there is a close correspondance between the mean predicted warming trend and the true trend across simulations (green dots in Figure S2a). The cases in which the approach is not able to capture the simulated trends as well correspond to those simulations which warm the most strongly and least strongly. This is to be expected, since any weighted average of warming trends across remaining models is expected to underestimate the warming trend magnitude in simulations from the model with the largest and smallest trends.

We also consider the imperfect model test using a metric based on RMSD of gridded SAT (blue dots in Figure S3) and a compound metric combining GSAT trend and RMSD of gridded SAT (red dots in Figure S3). The compound metric and RMSD-based metric show improved averages relative to unweighted results (based on correlation and RMSE; Figure S3 ; Table S4). However, the improvement using RMSD of SAT is largely a consequence of different models from the same institutions having similar SAT climatologies. We demonstrate this fact using an imperfect model test (Table S5) with one ensemble member per model for 2041-2060 (SSP5-8.5). When one of a pair of models from the same institution with very close RMSD of gridded GSAT (Figure S1a) is used as the pseudo-observations, the other will be assigned relatively heavy weight . When we remove one model from each pair from the same instution, the correlation coefficient for the metric RMSD of gridded GSAT declines from 0.32 (P =

0.09) to 0.21 (P = 0.32). Such a decrease does not occur for the metric based on GSAT trend (Table S5).

Text S7 Evaluation by probabilistic validation

a. Probabilistic validation using one randomly-selected ensemble member per model

For probabilistic validation of results based on random single-member per ensemble samples, we first apply the imperfect model test for each selection. We noted in which quintile of the projection (0-20%, 20-40%, etc) pseudo-observations lie for each projection, across all models for each selection and repeat the random selection 5000 times. The green and black bars in Figure S4 show the median of the 5000 single-member per model samples. The error bars show the corresponding +/-1 standard deviation ranges of each quintile.

b. Probabilistic validation using the ensemble mean for each model.

Probabilistic validation of weighting using the ensemble mean for each model is also based on the imperfect model test (see description as Text S6). In a imperfect model test, again, we noted in which quintile of the projection (0-20%, 20-40%, etc) pseudo-observations lie for each projection, across all models. Relative frequency calculations are scaled to account for differences in ensemble size between models [eg. when one member of a M-member ensemble acts as pseudo observation, the count number in the coresponding quintile will increase by 1/M]. In order to increase the sample of predictions for probabilistic validation, we note that in the 50-member ensemble of CanESM5, no correlation ($r = -0.11$, $p = 0.47$) is found between the 1970-2014 trend and the 2081-2100 vs 1995-2014 warming across this single-model ensemble. Internal variability includes a range of interannual and interdecadal processes (e.g. ENSO, AMO, and the PDO), which will not be correlated between different ensemble members. As a result, the internal variability in the earlier period is uncorrelated with that in the later period. Each realisation of projected 21st century warming from a given model is therefore equally probable given each realisation of its 1970-2014 trend, and therefore we use every possible combination of historical trend and future warming change for each model. This approach substantially increases the sample size, which should help with the probabilistic validation (Daines et al. 2016). Figure S5a and Figure S5b show the results of the probabilistic validation applied to mid-century and end-of-century projections respectively under SSP5-8.5. The frequency with which the pseudo-observations lie in each quintile of the unweighted Cumulative Distribution Function (CDF) is shown in black, and the frequency with which the pseudo-observations lie in each quintile of the narrower weighted CDF is shown in green. Overall the weighting approach gives approximately equal relative frequencies in each quintile, similar to the unweighted prediction.

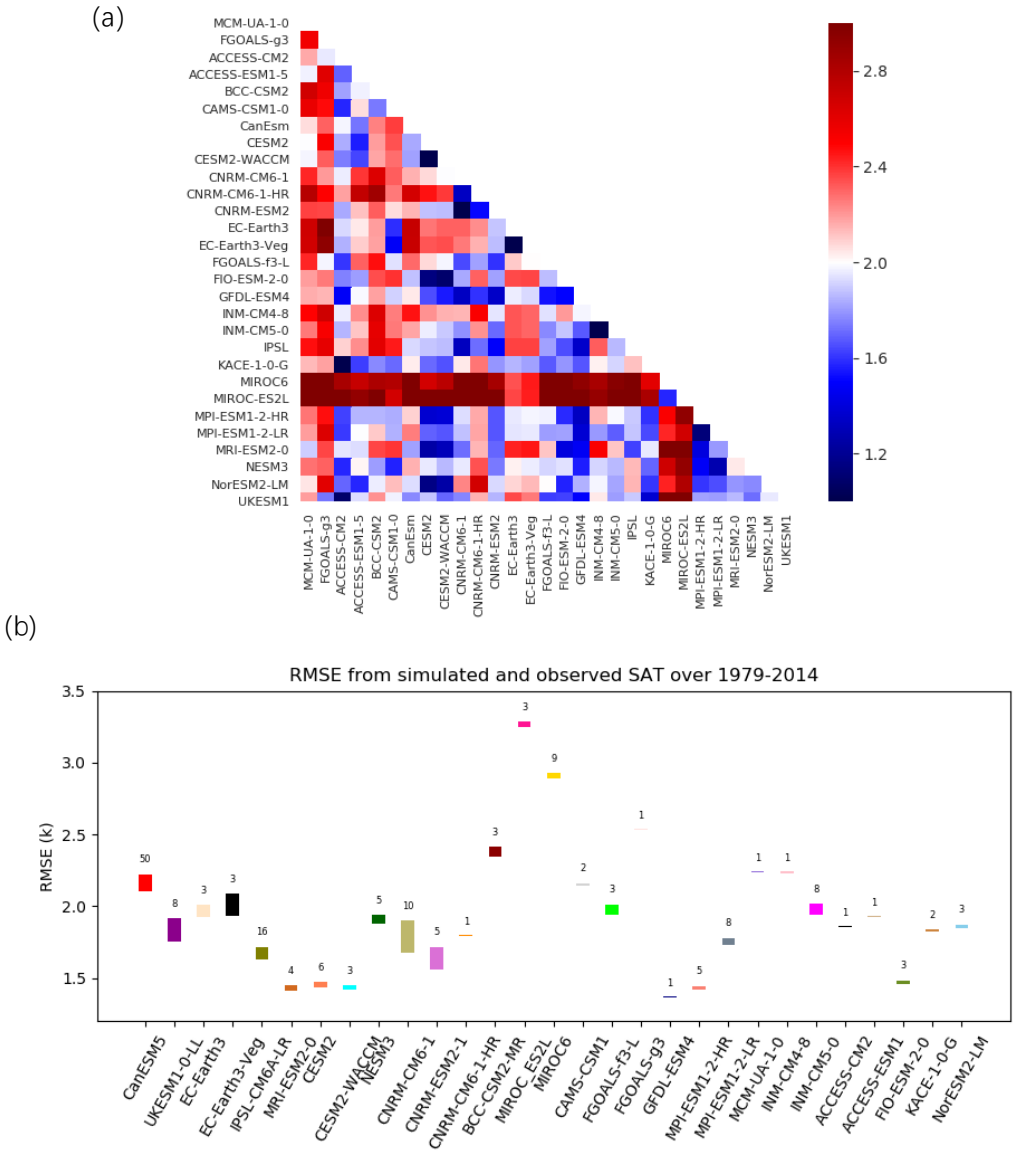


Figure AA.S1 (a) Inter-model RMSD in 1979-2014 mean gridded SAT (units: K). Each row and column represents a single climate model. Warm colors represent larger distance, while closer distances shown by cool colors. (b) Model and observation distance (units: K). Each bar represents a single climate model, the length of bar depend on the members' range. Models with large RMSD means far distance to observation, vice versa. The numbers marked at the bottom of each bar for panel b represent number of member in each model.

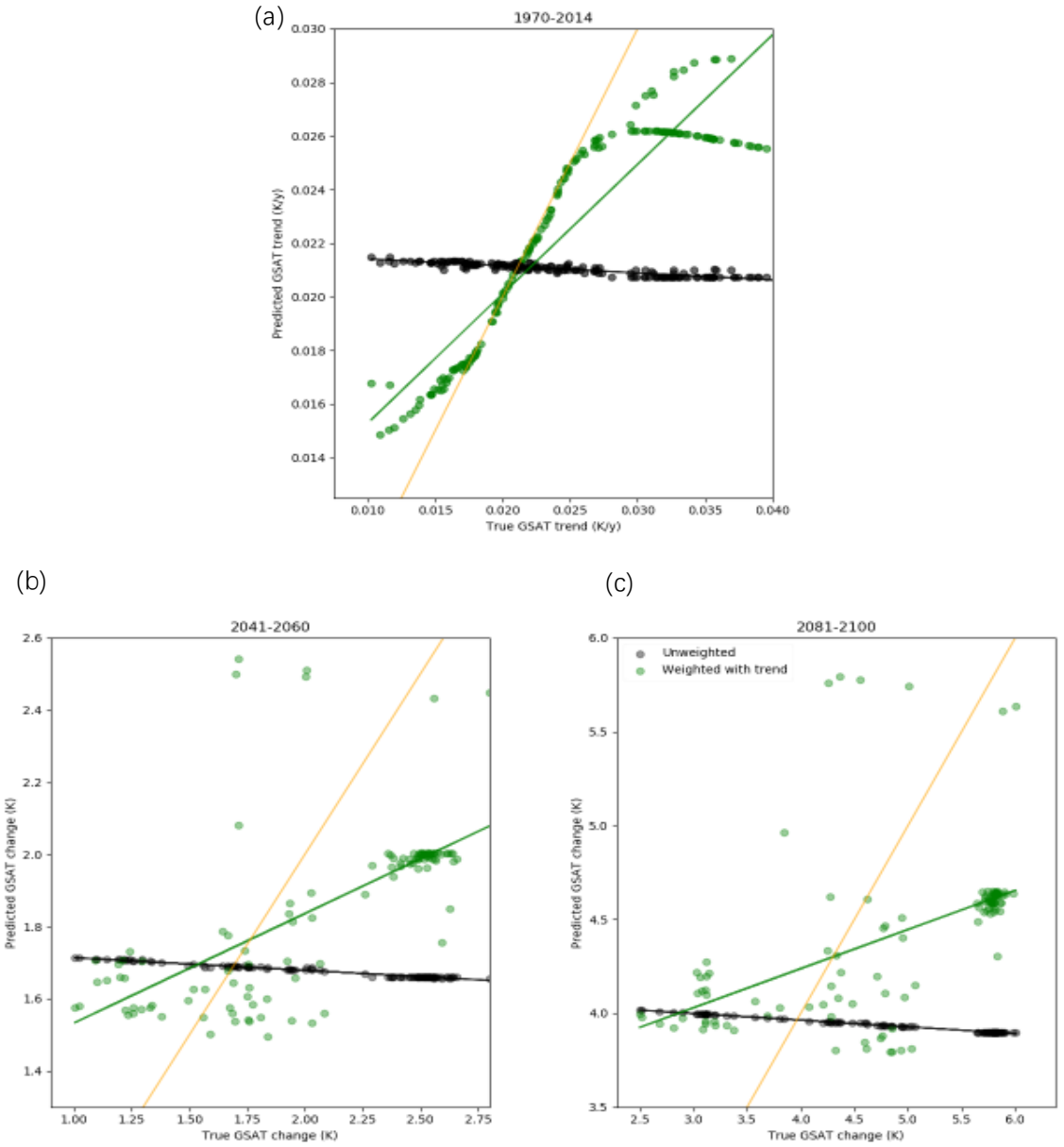


Figure AA.S2 Imperfect model test using ensemble means compared with the unweighted ensemble (green and black circles respectively) for historical (Panel a) and future periods of 2041-2060 (Panel b) and 2081-2100 (Panel c). In each plot, the x-axis represents pseudo observations and the y-axis represents the mean value predicted by the corresponding method. The orange line denotes the 1:1 line for which the predicted value is equal to pseudo-observations. The green and black lines are respectively linear least-squares fits for the weighting method and unweighted simulations. The values of correlation and RMSE for pseudo observations versus predicted means in this plot can be found in Table S4.

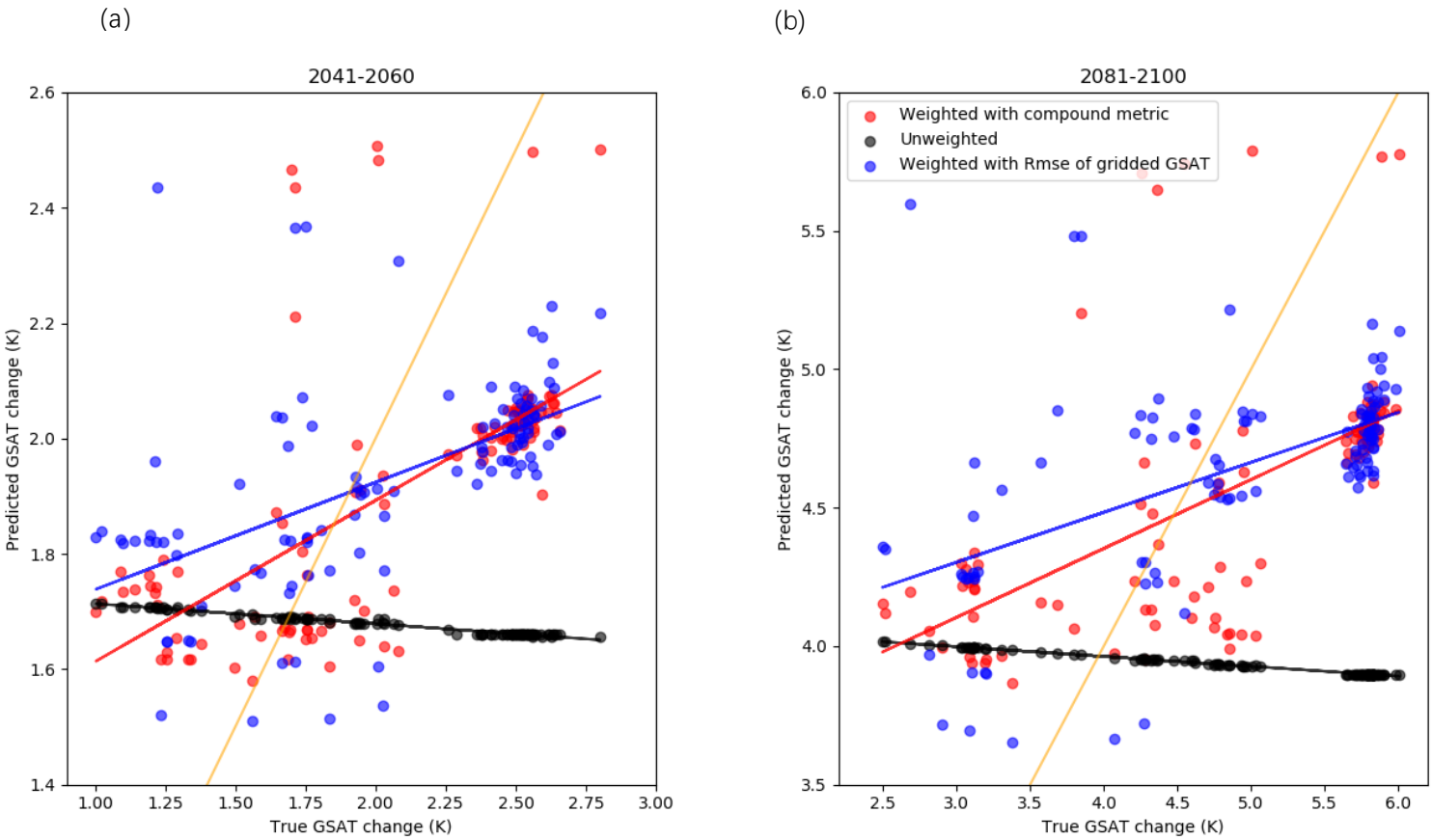


Figure AA.S3 Similar to Fig S2 but for the compound metric (red dots) and metric RMSD of gridded SAT (blue dots). The values of correlation and RMSE for pseudo observations versus predicted means in this plot can be found in Table S4.

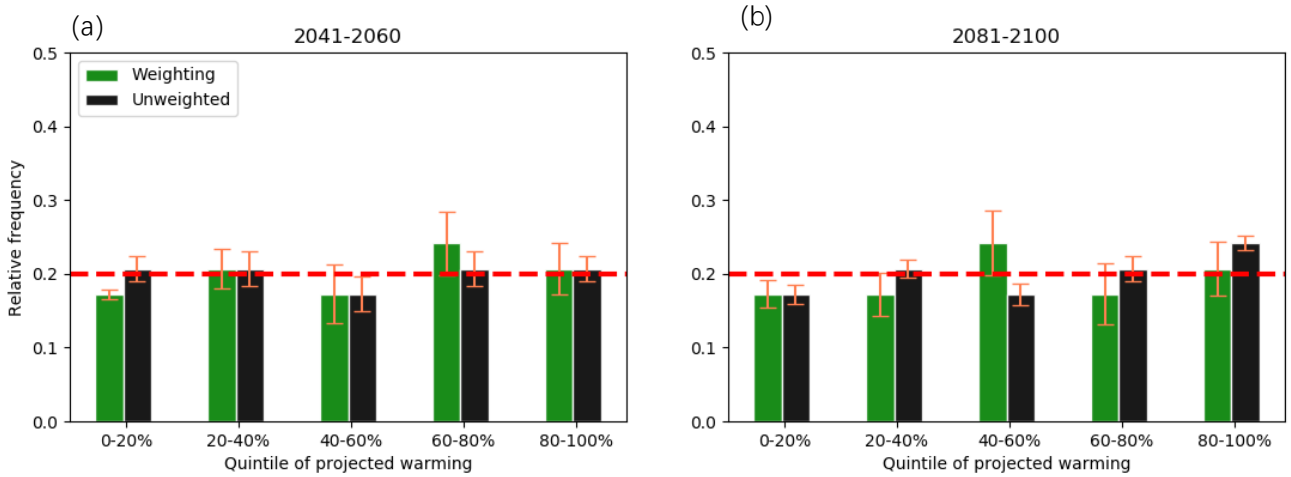


Figure AA.S4 Probabilistic validation using one randomly-selected ensemble member per model. Histograms show the relative frequency with which the true 21st century warming in the individual SSP5-8.5 simulation lies within each of five quintiles of projected warming derived using the weighting method and unweighted approaches in an imperfect model test, aggregated across all models. Bars denote the median of the

5000 single-member per model samples. Figure S4a and Figure S4b show results for 2041-2060 and 2081-2100 respectively, relative to the 1995-2014 base period. The green and black bars correspond to the weighting method and unweighted model output. The error bars show +/-1 standard deviation ranges for each quintile. Note that the 5th-95th percentile range of the weighted distribution is about 25% smaller than that of the unweighted distribution.

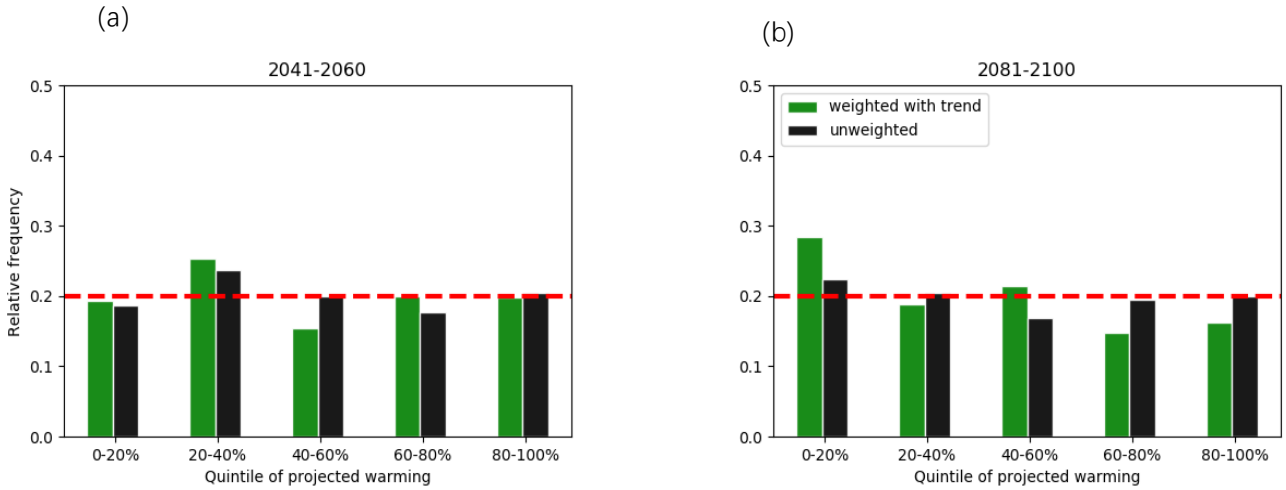


Figure AA.S5 Probabilistic validation using ensemble means for the weighting. Histograms show the relative frequency with which true 21st century warming in the individual SSP5-8.5 simulation lies within each of five quintiles of projected warming derived using the weighting method and unweighted approaches in an imperfect model test, aggregated across all models. Relative frequency is weighted such that each model has equal weight irrespective of ensemble size. Figure S5a and Figure S5b show results for 2041-2060 and 2081-2100 warming respectively, relative to the 1995-2014 base period. Note that the 5th-95th percentile range of the weighted distribution is about 25% smaller than that of the unweighted distribution.

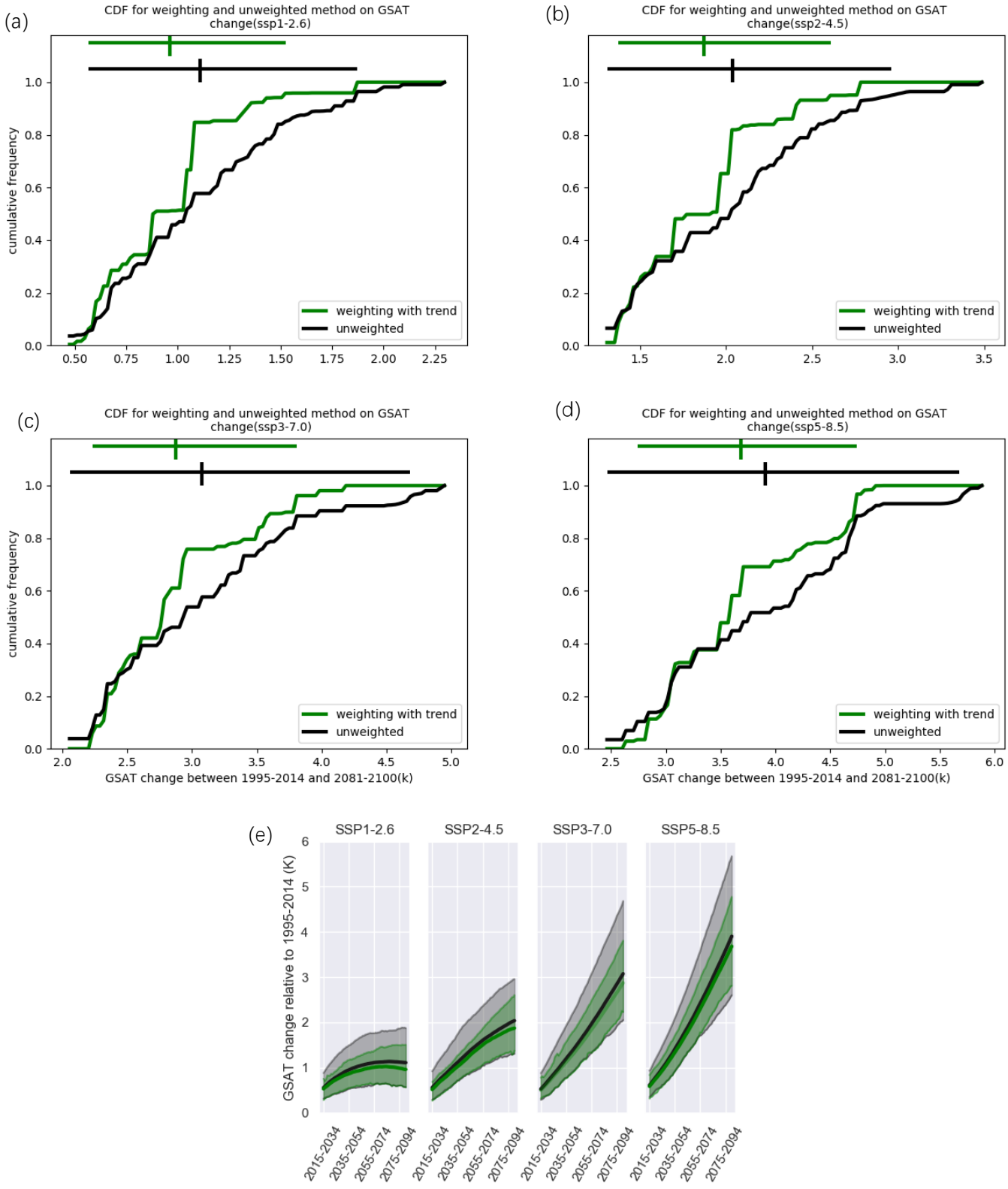


Figure AA.S6 Distributions of projected GSAT warming between 1995-2014 and 2081-2100 in each of four scenarios (Panel a-d), both constrained by observations (green) and unconstrained (black). Unconstrained projections are derived based on ensemble means of models, with equal weights given to each model. The weights for

the weighting method are calculated based on the corresponding ensemble mean of models' historical GSAT trends, and then give equal weights to ensemble members. Horizontal green and black lines show the corresponding 5-95% ranges, and the vertical ticks show the corresponding means. Panel e shows 5-95% ranges of weighting (green shadow) and unweighted results (grey shadow) for other projection periods. The green (black) solid line show the corresponding means of weighting (unweighted) results.

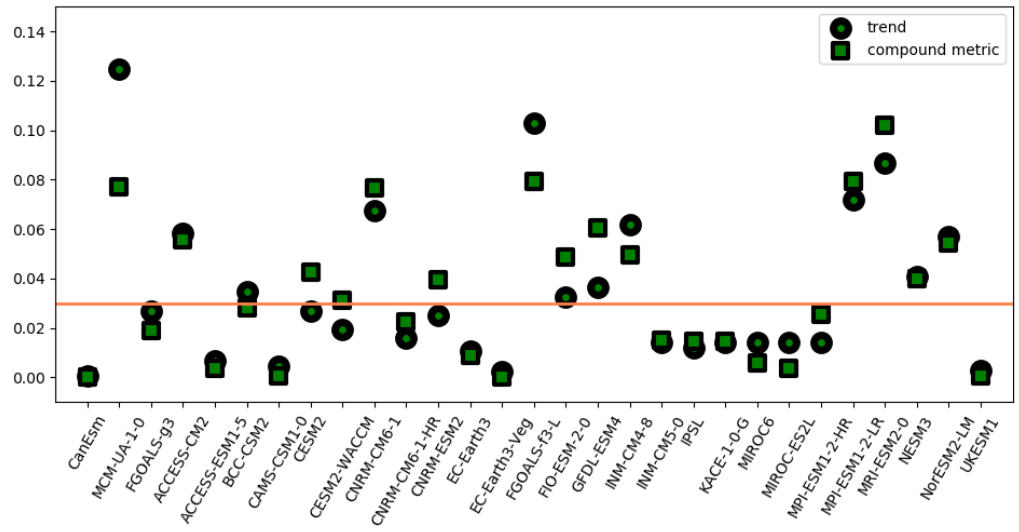


Figure AA.S7 The models weights obtained by metric GSAT trend (circles) and compound metric (squares) from SSP5-8.5. Model weights are calculated by ensemble mean of each model. The orange line represents equal weights.

Appendix B

2.1 Materials and methods

2.1.1 Linear regression method

The regression model used to implement the emergent constraints of future GSAT changes is

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} \tag{3}$$

where $\boldsymbol{y}$ is the vector of modelled future GSAT change (each element corresponding to a different model), the vector $\boldsymbol{\alpha}$ and the vector $\boldsymbol{\beta}$ are the multiple regression parameters, and the metrics form the matrix $\boldsymbol{X}$ ($\boldsymbol{X}^{\boldsymbol{T}}$ is the transpose of $\boldsymbol{X}$, the matrix which includes all metrics for all models. The number of rows in $\boldsymbol{X}^{\boldsymbol{T}}$ is the number of climate models we use; the number of columns is the number of metrics we select). We estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as $\hat{\alpha}$ and $\widehat{\boldsymbol{\beta}}$ using ordinary least squares regression, with $\boldsymbol{X}$ and $\mathbf{y}$ taken from CMIP6 models (using single randomly drawn members from each model with an initial condition ensemble; Section 3.2.4). Applying observational estimates $\boldsymbol{X_0}$ of the corresponding selected metrics (sampled from the observational uncertainty) with the linear regression model yields a multi-diagnostic constraint $\hat{y}_0$ (Karpechko et al. 2013; Senftleben et al. 2020):

$$\hat{y}_0 = \hat{\alpha} + \boldsymbol{X_0^T}\widehat{\boldsymbol{\beta}} \tag{4}$$

If we assume Gaussian residuals as in previous studies (Schlund et al. 2020; Tokarska et al. 2020), the probability density function (PDF) for projected GSAT changes (y), given observation, $\boldsymbol{X_0}$ is

$$p(y|\boldsymbol{X_0}) = \frac{1}{\sqrt{2\pi\sigma_{\hat{y}_0}^2}} \exp\left(-\frac{(y-\hat{y}_0)^2}{2\,\sigma_{\hat{y}_0}^2}\right) \tag{5}$$

where

$$\sigma_{\hat{y}_0}^2 = s^2(1 + \boldsymbol{X_0^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0}) \tag{6}$$

and

$$s^2 = \frac{1}{n-p-1}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2 \tag{7}$$

The derivation of Eqn.(6) is provided in Hooper and Zellner (1961). In Eqn.(7), M is the number of models and p denotes the number of metrics included in the regression; n is equal to the value of 26 when we take the full number of models, otherwise n is equal to the value of 20 by accounting for model dependence (Appendix 2.1.3).

As described in section 3.2.4, we assess the effect of internal variability by taking one random selection of one ensemble member per model and simultaneously drawing one

random choice of each observed metric from our estimate of the observed distribution. We repeat this process 10000 times, and derive 10000 estimates of the PDF of projected warming (shown as shading in Fig 3.10). We next demonstrate that the sample mean of these PDFs (shown as solid lines in Fig 3.10) is an estimate of the population estimate of the PDF of projected warming.

When our study samples over initial condition ensembles, each sample gives us a distribution of projected GSAT change (noted as x in equation 8) with a particular mean (μ) and standard deviation (σ). Hence, multiplying the joint distribution of these two statistics $f(\mu, \sigma)$, by a conditional distribution $f(x|\mu, \sigma)$, and then integrating over μ and σ, gives us a population estimate $p(x)$ of the marginal pdf of projected GSAT changes.

$$p(x) = \iint f(x|\mu, \sigma)\, f(\mu, \sigma)\, d\mu\, d\sigma \tag{8}$$

Based on our sampling strategy, each derived PDF with its corresponding value of μ and σ is equally probable. Sampling μ and σ from their joint distribution and then averaging the resulting conditional distributions of x gives a sample estimate of this population mean distribution (solid curves of constrained PDFs).

2.1.2 Weighting method

To relate our results to those of Chapter 2, we consider a weighting method based on (Sanderson et al. 2015a, 2015b) and Knutti et al. (2017) as an alternative approach for deriving observationally constrained projections. We first define

$$w_i = \frac{e^{-\frac{Di^2}{\sigma_D^2}}}{1 + \sum_{j \neq i}^{M} e^{-\frac{{S_{ij}}^2}{{\sigma_s}^2}}} \tag{9}$$

We then obtain ***W,*** a vector of weights for each model, by normalizing w_i by the sum of w_i. In equation (9), M is the number of models considered, D_i is the difference between the simulated value of the metric from model i and the observed value, and S_{ij} is the difference between models i and j for the selected metric. When we weight models with multiple metrics, we calculate D_i and S_{ij} giving equal weights to each metric after normalizing each metric by the median across models. The parameter σ_S regulates the degree to which model similarities are penalized in the weights, and σ_D regulates the effect of model performance on the weights (Knutti et al. 2017; Lorenz et al. 2018). We calculate σ_S and σ_D based on approaches proposed by Knutti et al. (2017), Lorenz et al. (2018) and Brunner et al. (2019a) (Text S1 and S2). In contrast to the linear regression approach, this approach weights model projections based not only on their goodness of fit to historical observations but also based on similarity between models. The model weights obtained using Eq (9) not only yield the constrained multi-model mean, but also are used to weight models to obtain a multi-model distribution for the constrained uncertainty range (Section 3.2.5).

2.1.3 Model dependence

Models from different institutions may share model components, and some individual modelling centres produce multiple closely related model versions. As a result, the individual members of the multimodel ensemble are not expected to be independent. This interdependence is accounted for in the weighting approach by downweighting models that are similar in their historical simulations. For the linear regression approach, the number of independent models must be estimated in order to determine the number degrees of freedom in the regression. To this end, we assess model dependence using the "model genealogy" method (Masson; Knutti 2011a). To build up the CMIP6 model genealogy, we apply a hierarchical clustering method (Brunner et al. 2020a; Knutti et al. 2013; Masson; Knutti 2011a) implemented by the linkage function in Python SciPy. The intermodel distance is defined as the area weighted mean square difference in near-surface air temperature averaged over 1970-2014 on a 1° grid between two models. We determine the number of independent models by comparing inter-model distances with inter-ensemble member distances in a model genealogy approach (Text S1, Fig S1 and Fig S2). Based on the results of this analysis, in Section 3.3.2 – 3.3.4 the number of independent models considered in the linear regression approach is set to 20 rather than the full number of models. The sensitivity of the stepwise selection procedure to the number of statistical degrees of freedom used is presented in Fig S3 and S4.

2.2 Supplementary information

Table AB.2. S 1 As Table 3.2, but derived using a value for the number of statistical degrees of freedom equal to the total number of models.

Metrics	Projected warming (5-95%, units: K)	
	SSP5-8.5	SSP1-2.6
Unconstrained	4.08 (2.53, 5.62)	1.30 (0.46, 1.95)
GT	3.76 (2.59,4.93)	1.08 (0.42,1.75)
MBLC+BCS	3.99 (3.01,5.02)	1.15 (0.67,1.62)

Text S1 Calculation of independence weighting parameter and number of statistical degrees of freedom

The parameter σ_s in the weighting approach controls the importance of model similarity in the weighting. We estimate σ_s for the independence weight ($1 + \sum_{j \neq i}^{M} e^{-\frac{S_{ij}{}^2}{\sigma_s{}^2}}$; the denominator of eq 9 in Appendix 2.1.2) by a method proposed by (Brunner et al. 2019b) [more details in Text S1 of Chapter 2]. We then calculate squared differences across models for the metrics considered and then normalize these by their respective medians. We then give equal weights to each metric to obtain S_{ij}. Applying the method of Brunner et al. (2019), we obtain σ_s= 0.61 for the weighting approach including MBLC and GT metrics, and σ_s = 0.82 for the weighting approach including MBLC and BCS metrics. We can use σ_s as a threshold to determine the number of independent models based on a model genealogy. Since the model genealogy is based on pairs of model-model distance calculated using a spatially-distributed variable, we calculate a model-model distance matrix using the climatological near-surface air temperature field over the 1970-2014 period (Fig S1). We then calculate σ_s (σs=0.58) as described in Brunner et al. (2019a) based on gridded near surface air temperature. Using the model-model distance matrix (Fig S1), we then build up the model genealogy (details described in Appendix 2.1.3), as shown in Fig S2. Using this result we estimate a value of 20 for the number of statistical degrees of freedom in the CMIP6 ensemble.

Text S2 Weighting method with selected constraints

Step 1: We randomly pick one ensemble member per model for the historical simulation and pick the same ensemble member for the SSP scenario simulation.

Step 2: We then apply an imperfect model test to determine σ_D from Eq 9 in the manuscript using the historical simulation of GSAT changes. As described in previous studies (Brunner et al. 2019a; Knutti et al. 2017; Lorenz et al. 2018), we expect 90% of the pseudo-observations to lie within the 5-95% projected range for the selected σ_D [more details in Text S2 of Chapter 2].

Step 3: We apply the observed quantities to the weighting approach (Eq 9 in

manuscript) using historical simulations to weight models, and then apply the weights to projected GSAT changes from each model to estimate projected ranges and the weighted mean.

Step 4: We repeat Steps 1-3 10,000 times.

Text S3 Sampling over internal variability.

Since 14 of 26 models only have one ensemble member, the process of random selection of one ensemble per model (described in Section 3.2.4) can only partially account for internal variability.

To investigate how well our random selection approach can estimate the distribution of internal variability, we carry out the following analysis:

Step 1: We first artificially set all ensemble sizes to one (through random selection of individual realizations from models providing more than one ensemble member).

Step 2: We randomly replace a number M of the models with all available ensemble members (varying M=1,2,3…12 in turn) and repeat the sampling calculation described in Section 3.2.4 to account for internal variability. We apply the observational constraint with the best estimate of the observed quantity described in Appendix 1.1 Eq (5) to see if the width of the predicted PDFs of 5^{th} and 95^{th} percentiles of GSAT changes converges by the time we have included all available ensembles.

Step3: We then repeat Step 2 1000 times and get the mean uncertainty of the sampling range for each M. As shown in Fig S5, the uncertainty contributed by internal variability is still increasing with M (especially the 95^{th} percentiles of ensemble spread) for metrics involving the GSAT trend. This result implies that the effects of internal variability are underestimated compared to the case where multi-member ensembles were available for all models. However, this is not the case for cloud metrics which remain stable regardless of how many model ensemble members are chosen.

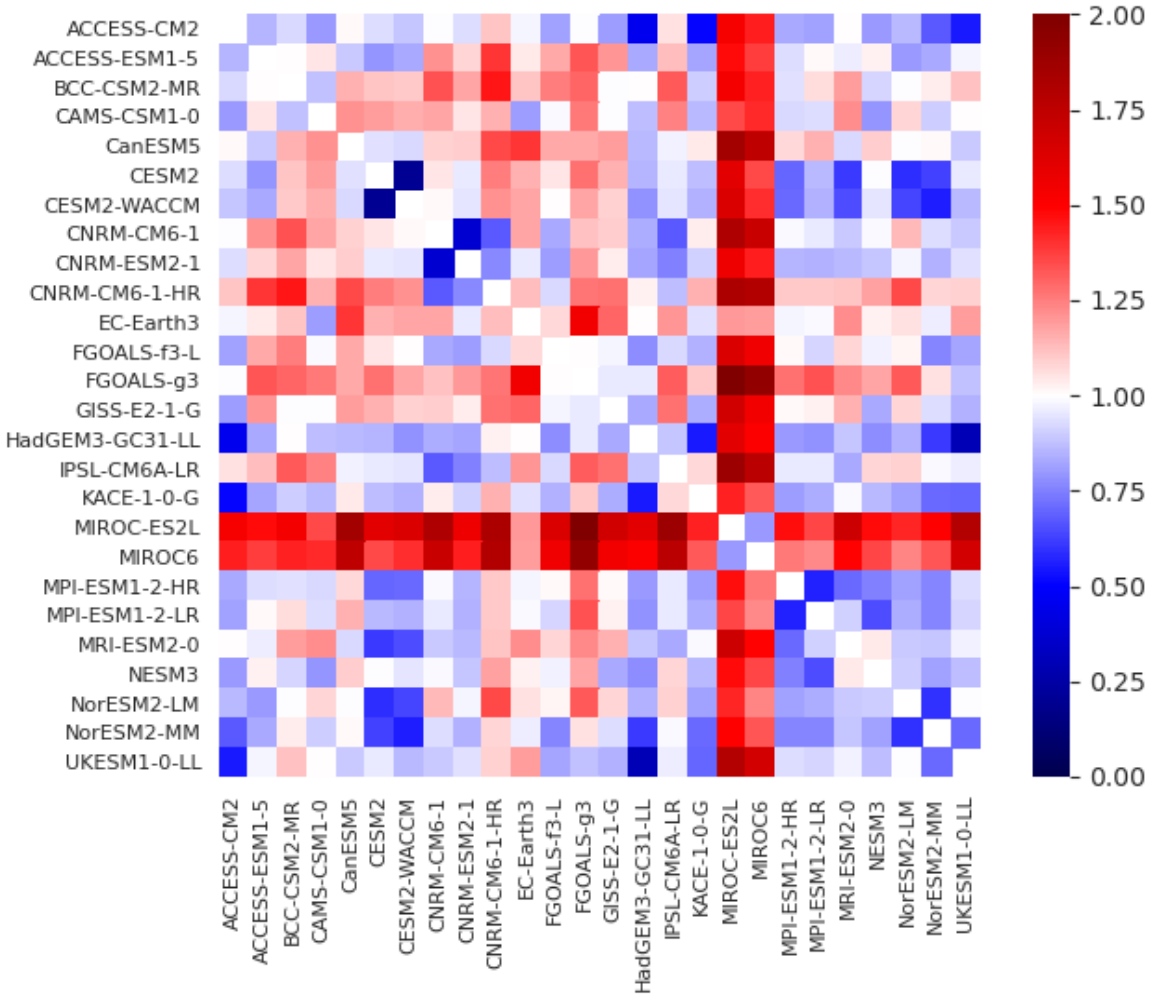


Figure AB.S1 Model-model distance matrix normalized by its median for gridded annual mean surface air temperature over the period of 1970-2014 (see details in Text S2 and Appendix 3.1.3).

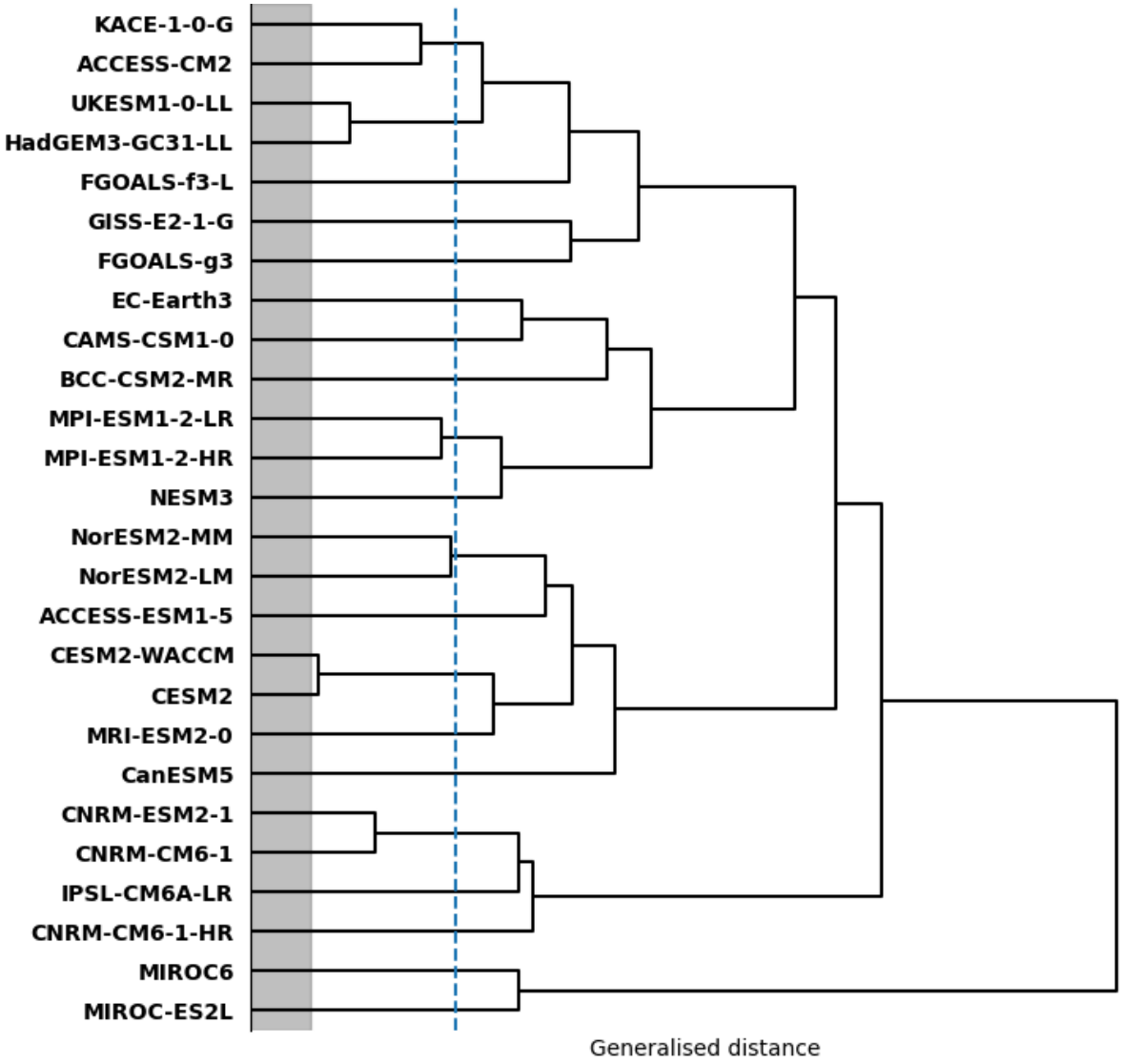


Figure AB.S2 Model family tree for all 26 CMIP6 models used in this study. Models branching further to the right are more independent. Fig S2 is based on the model-model distance matrix shown in Fig S1 calculated from gridded near-surface air temperature. The dashed vertical line represents the independence shape parameter and is used to determine models that are independent or not (Brunner et al. 2020b). The gray shading represents an estimate of internal variability calculated from the median of distance between pairs of initial-condition realizations taken from the same model.

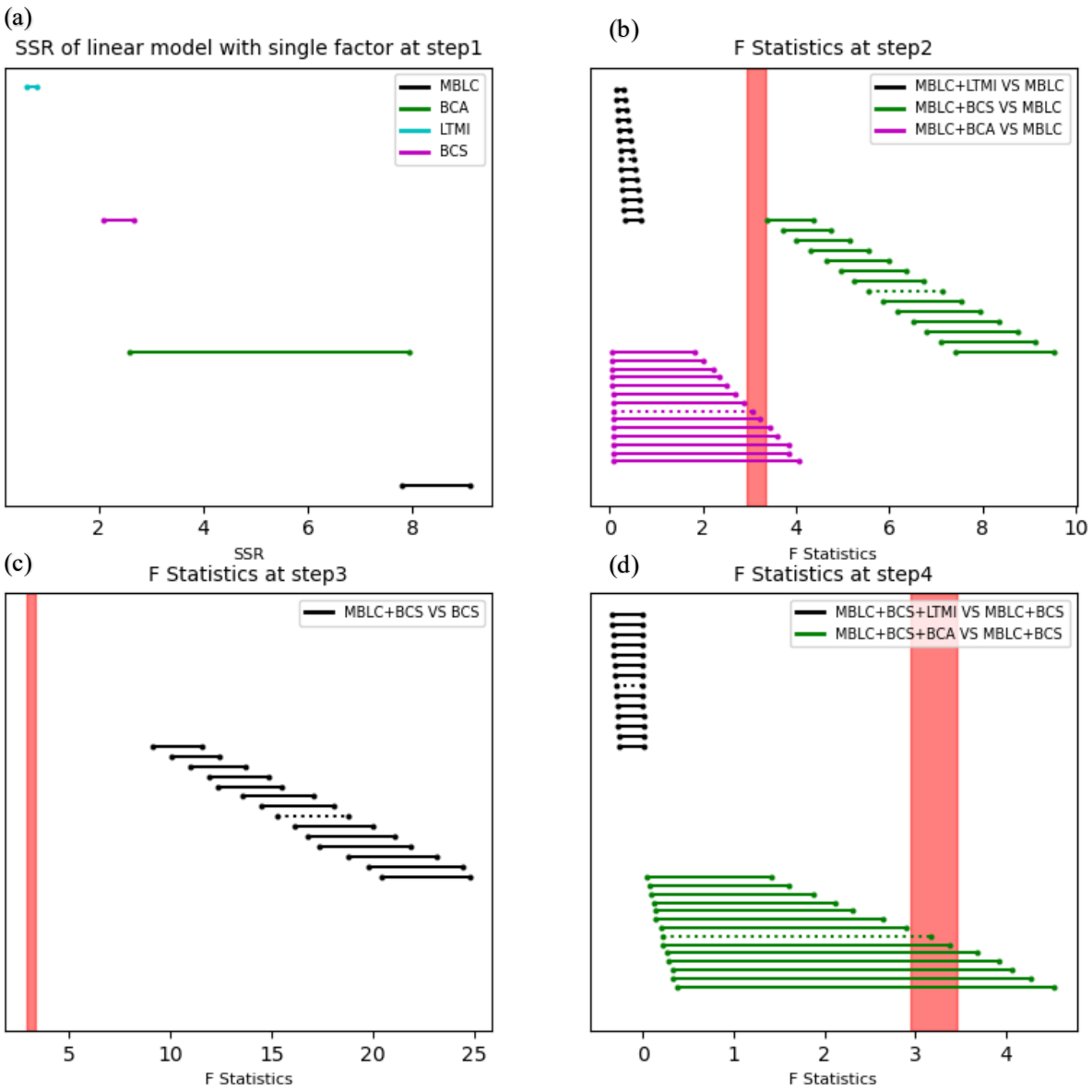


Figure AB.S3 Similar to Fig 3.3 but for F statistics calculated with a range of numbers of degrees of freedom. For each class of F statistic in the same color bar, the bars correspond to numbers of degrees of freedom ranging from the full number of models (topmost bar) to half the number of models (bottommost bar). The horizontal dashed lines represent the F statistics with effective degree of freedom determined from the models' genealogy (Appendix 3.1.3, Text S2), as also displayed in Figure 3.3. The vertical red shaded areas represent critical F values at the 0.1 level for the corresponding range of degrees of freedom.

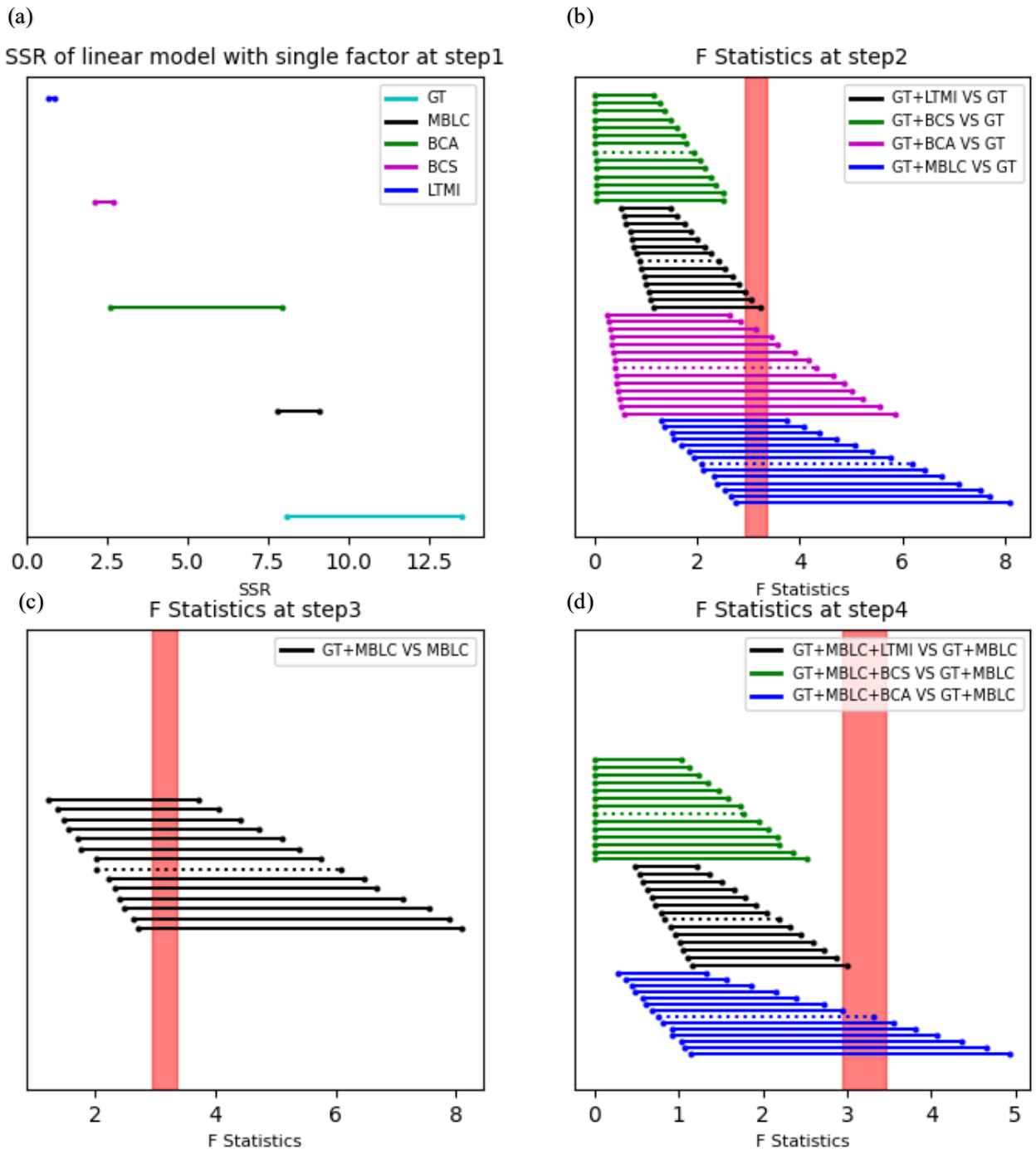


Figure AB.S4 As in Fig S3 but with the stepwise selection based on low cloud and GSAT trend metrics.

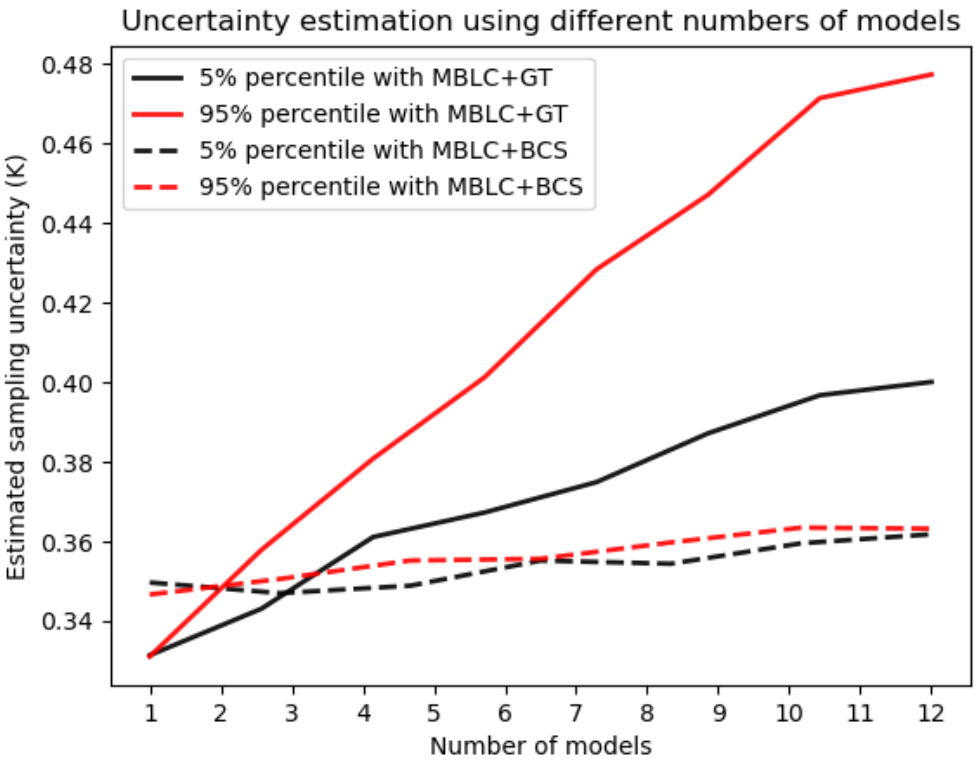


Figure AB.S5 An assessment of how well the effects of internal variability are accounted for in our study, as described in Text S4. We first artificially set all ensemble sizes to one. We then select M models, with, M=1,2,3…12 in turn, from the set of models with ensemble size greater than one, and repeat the sampling calculation (described in Section 3.2.4) to get the distribution width (maximum minus minimum) of observationally constrained 5th (in black) and 95th (in red) percentiles. For each number of models, we show the mean across the sampling range. Solid lines correspond to MBLC and GSAT trend metrics, dashed lines to MBLC and BCS metrics.

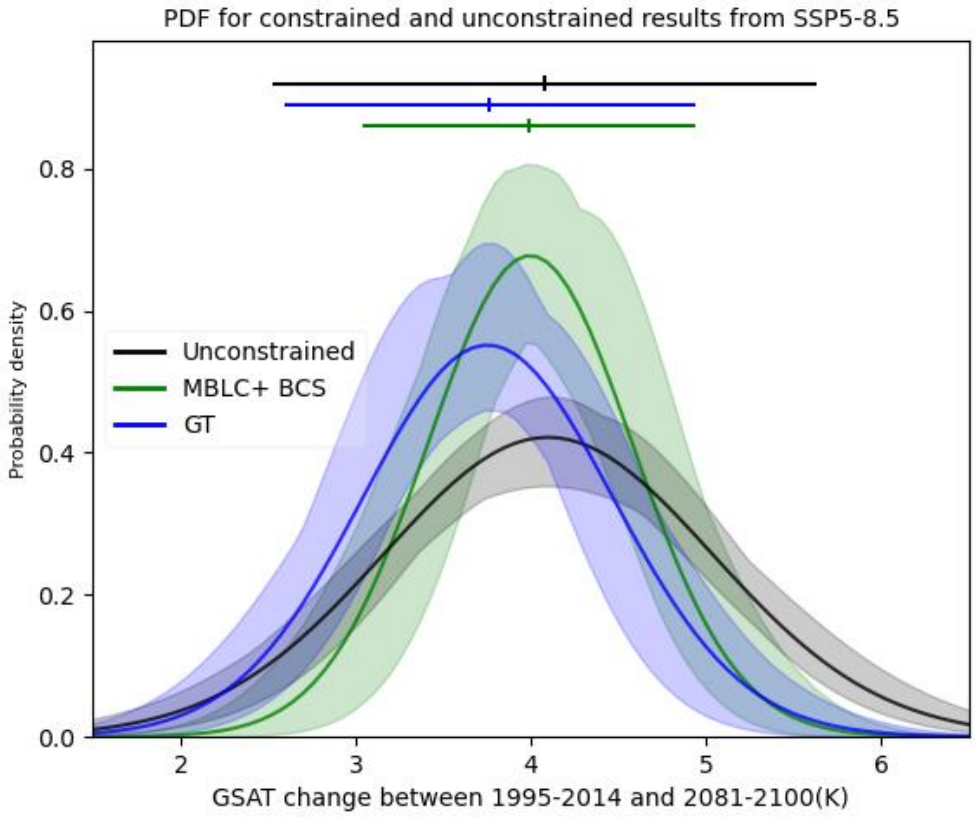


Figure AB.S6 Similar to Fig 3.9, except using a number of statistical degrees of freedom equal to the number of models in the CMIP6 ensemble.

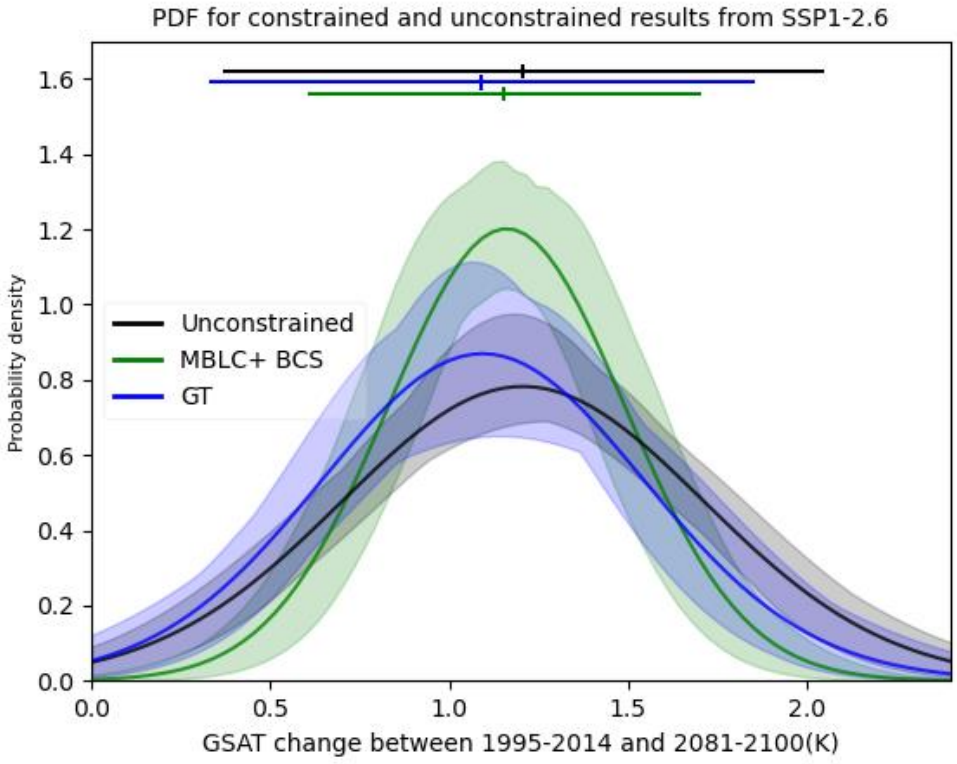


Figure AB.S7 Similar to Fig3. 9, but for SSP 1-2.6.

Appendix C

3.1 Materials and methods

3.1.1 Metric selection strategy

3.1.1.1 Stepwise selection process

One approach to selecting regional metrics applies a stepwise selection process to each region separately. The pool of potential metrics for the stepwise selection process is shown in Table 4.2, and includes global and regional metrics together. For each of the regional metrics we consider the trend of time series, the climatology (mean of time series) and the variability (standard deviation of detrended time series). We use ensemble means of potential metrics and regional projected warming for models that have more than one ensemble member to apply the stepwise selection process. We apply a standard process of forward stepwise selection using the Akaike Information Criterion [AIC] to build up a multivariate regression model that can be the best model with combination of predictability and robustness for the future regional warming. The following steps are used:

(1) We first approximate the response variable y of regional warming with a constant by building up a regression model with an intercept only.
(2) We then successively add individual metrics into the regression model.
(3) At each step, the added metric is the one that best improves the accuracy (measured by the AIC) in prediction of the response variable y.

In summary, the 'best' one can be chosen which has the lowest AIC from a selection of candidate models. The step wise selection process will terminate when the AIC increases with the addition of more metrics.

3.1.1.2 Lasso (least absolute shrinkage and selection operator) regression approach

Based on ordinary least squares (OLS), the Lasso approach (eq 7) adds a penalty term to the residual sum of squares (RSS). The Lasso approach applies the multiple linear regression approach used in the main study, but removes metrics to minimize a cost function which is represented here

$$\widehat{\boldsymbol{\beta}} = arg_{\widehat{\boldsymbol{\beta}}} min\ [\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\right)^2 + \lambda \sum_{i=1}^{k}|\beta_i|] \quad (7)$$

In eq (7), n represents sample size, which is the number of models in this study, while k represents the number of regression coefficients corresponding to the number of metrics. y_i is the modelled future regional warming change (each i corresponds to a

different CMIP6 model), and the metrics form the $\boldsymbol{x}_i$, the vector which includes all metrics. $\widehat{\boldsymbol{\beta}}$ is the vector of regression coefficients determined by eq (7). $\hat{\beta}_0$ represents an intercept term in the regression model. With optimal λ, some of the β_i coefficients are reduced to 0. In this way, the Lasso serves as a metric selection approach.

A range of λ are tested to select the optimal λ using cross-validation. The following steps are involved in the cross-validation for a given value of λ:

- Separation of data into a test and a training set
- Estimating regression coefficients with models from the training set
- Obtaining predicted values in the test set using the regression coefficients estimated from the training set
- Calculating mean squared prediction error (MSPE) between predicted values and true values in the test set

This study applies *k*-fold (in our study, *k*=5) cross-validation. In the 5-fold cross-validation, the data are separated into equal parts where each part will serve as the test set (one of the 5-fold) in turn and the rest of data is training set. We average MSPE across all the folds and end up with the optimal λ with a lowest average MSPE.

3.1.2 Weighting approaches

We also apply the Sanderson weighting approach to constrain projections (Sanderson et al. 2015a, 2015b) . We define

$$w_i = \frac{e^{-\frac{Di^2}{\sigma_D^2}}}{1 + \sum_{j \neq i}^{M} e^{-\frac{{S_{ij}}^2}{{\sigma_s}^2}}} \tag{8}$$

In equation (8), M denotes to the number of models in use, D_i quantifies the difference between the observation and the model simulation of the metric from model *i*, and S_{ij} quantifies the difference between models *i* and *j* for the metric considered. σ_D and σ_s are the parameters determining the effect of model performance and the degree of model similarities on the model weights respectively (Knutti et al. 2017; Lorenz et al. 2018). We estimate σ_s and σ_D using the approach described in previous studies (Brunner et al. 2019b; Brunner et al. 2020a; Brunner et al. 2020c; Knutti et al. 2017; Lorenz et al. 2018). Based on Eq (8), by normalizing w_i by the sum of w_i, we can obtain a vector of weights $\boldsymbol{W}$, containing weights for each model. More details are presented in Chapter 3.

3.2 Supplementary information

Text S1.

We carry out a synthetic data experiment to test whether our uncertainties are underestimated due to the availability of only a single ensemble member from many

CMIP6 models. The steps of our synthetic data experiment are described in the following paragraphs.

Step1: Build 50-member synthetic ensembles of the observable metrics and projected warming for each model, centered on the corresponding means for each model, and sampled from a Gaussian distribution with standard deviation equal to that of the CanESM5 large ensemble.

Step 2: Sample one ensemble member per model from these synthetic 50-member ensembles and generate a PDF. Repeat this process 5000 times, and average (the same sampling approach as used in Section 4.2d). The resulting averaged PDF is shown as the red solid PDF in Fig S1.

Step 3: Drawing from the 50-member ensembles, pick the same number of ensemble members for each model as are actually available (Table S1) and use them in the remainder of this step. Apply the same sampling method as in Step2 to sample individual ensemble members from this subset and generate an average PDF (blue dashed lines).

Step 4: Repeat Step 3 using a different subset of ensemble members.

Our results in Fig S1 show that the red dashed line is close to the blue dashed lines, suggesting that the limited ensemble sizes available do not have much effect on our resulting PDF, using MBLC as a predictor for projected warming in the Northern West North America (NWN) region. We also applied the Step 1 to Step 5 for other regions and similar conclusions were obtained.

Table AC.2. S 1 CMIP6 Historical, SSP1-2.6 and SSP5-8.5 simulations used in this study. The number of ensemble members provided for each forcing senario is indicated in the second through the fourth columns.

Model name	**Historical**	**SSP1-2.6**	**SSP5-8.5**
ACCESS-CM2	1	1	1
ACCESS-ESM1	1	1	1
BCC-CSM2-MR	3	1	1
CAMS-CSM1-0	2	2	2
CanESM5	50	50	50
CESM2	6	1	2
CESM2-WACCM	3	1	1
CNRM-CM6-1	10	6	6
CNRM-CM6-1-HR	1	1	1
CNRM-ESM2-1	5	5	5
EC-Earth3	3	3	3
FGOALS-f3-L	3	1	1
FGOALS-g3	1	1	1

GISS-E2-1-G	1	1	1
HadGEM3-GC31-LL	4	1	1
IPSL-CM6A-LR	16	3	5
KACE-1-0-G	3	2	2
MIROC-ES2L	3	1	1
MIROC6	9	3	3
MPI-ESM1-2-HR	5	1	1
MPI-ESM1-2-LR	8	8	8
MRI-ESM2-0	4	1	1
NESM3	5	2	2
NorESM2-LM	3	1	1
NorESM2-MM	3	1	1
UKESM1-0-LL	8	5	4

Table AC.2. S 2 Projected mean warming and 5-95% confidence ranges based on the constrained and unconstrained projections for two SSP scenarios over 2081-2100 relative to 1995-2014 (units: K).

Scenarios	Regions	Mean and 5–95th confidence ranges		Regions	Mean and 5–95th confidence ranges	
		Unconstrained	Constrained		Unconstrained	Constrained
SSP1-2.6	**GIC**	1.80 (-0.74, 4.34)	1.50 (-0.46, 3.48)	**NWN**	2.51 (0.39, 4.63)	2.40 (0.88, 3.92)
SSP5-8.5		6.40 (2.88, 9.90)	6.08 (3.52, 8.63)		7.57 (3.92, 11.20)	7.35 (4.99, 9.70)
SSP1-2.6	**NEN**	2.61 (0.36, 4.85)	2.48 (1.01, 3.96)	**WNA**	1.71 (0.49, 2.92)	1.66 (0.80, 2.52)
SSP5-8.5		8.53 (4.64, 12.41)	8.38 (5.92, 10.84)		5.43 (3.31, 7.54)	5.31 (3.92, 6.71)
SSP1-2.6	**CNA**	1.82 (0.56, 3.06)	1.80 (1.04, 2.55)	**ENA**	1.67 (0.58, 2.75)	1.68 (1.06, 2.29)
SSP5-8.5		5.83 (3.68, 7.98)	5.72 (4.27, 7.17)		5.24 (3.24, 7.22)	5.19 (3.87, 6.50)
SSP1-2.6	**NEU**	1.35 (-0.24, 2.95)	1.17 (-0.03, 2.39)	**WCE**	1.52 (0.17, 2.87)	1.45 (0.45, 2.43)
SSP5-8.5		4.46 (2.08, 6.85)	4.27 (2.48, 6.04)		5.17 (2.82, 7.51)	5.09 (3.64, 6.55)
SSP1-2.6	**EEU**	1.95 (0.41, 3.48)	1.85 (0.84, 2.87)	**MED**	1.40 (0.47, 2.32)	1.35 (0.66, 2.05)
SSP5-8.5		6.49 (3.65, 9.32)	6.39 (4.69, 8.10)		4.79 (3.19, 6.39)	4.71 (3.60, 5.81)
SSP1-2.6	**RAR**	3.21 (0.33, 5.94)	2.93 (1.03, 4.83)	**WSB**	2.10 (0.60, 3.58)	2.01 (1.06, 2.97)
SSP5-8.5		9.44 (5.09, 13.74)	9.13 (6.27, 12.00)		7.01 (3.99, 10.02)	6.88 (5.06, 8.72)
SSP1-2.6	**ESB**	2.02 (0.53, 3.49)	1.96 (0.88, 3.04)	**RFE**	2.36 (0.50, 4.20)	2.28 (0.89, 3.66)
SSP5-8.5		6.80 (3.87, 9.73)	6.70 (4.73, 8.65)		6.77 (3.68, 9.86)	6.63 (4.29, 8.98)
SSP1-2.6	**WCA**	1.63 (0.63, 2.63)	1.55 (0.89, 2.21)	**ECA**	1.69 (0.70, 2.68)	1.63 (0.87, 2.38)
SSP5-8.5		5.77 (3.60, 7.94)	5.64 (4.12, 7.15)		6.09 (3.69, 8.50)	5.92 (4.24, 7.60)
SSP1-2.6	**TIB**	1.64 (0.70, 2.57)	1.56 (0.79, 2.34)	**EAS**	1.61 (0.65, 2.57)	1.59 (0.95, 2.23)
SSP5-8.5		5.86 (3.51, 8.19)	5.66 (3.75, 7.57)		4.84 (2.97, 6.71)	4.77 (3.59, 45.95)

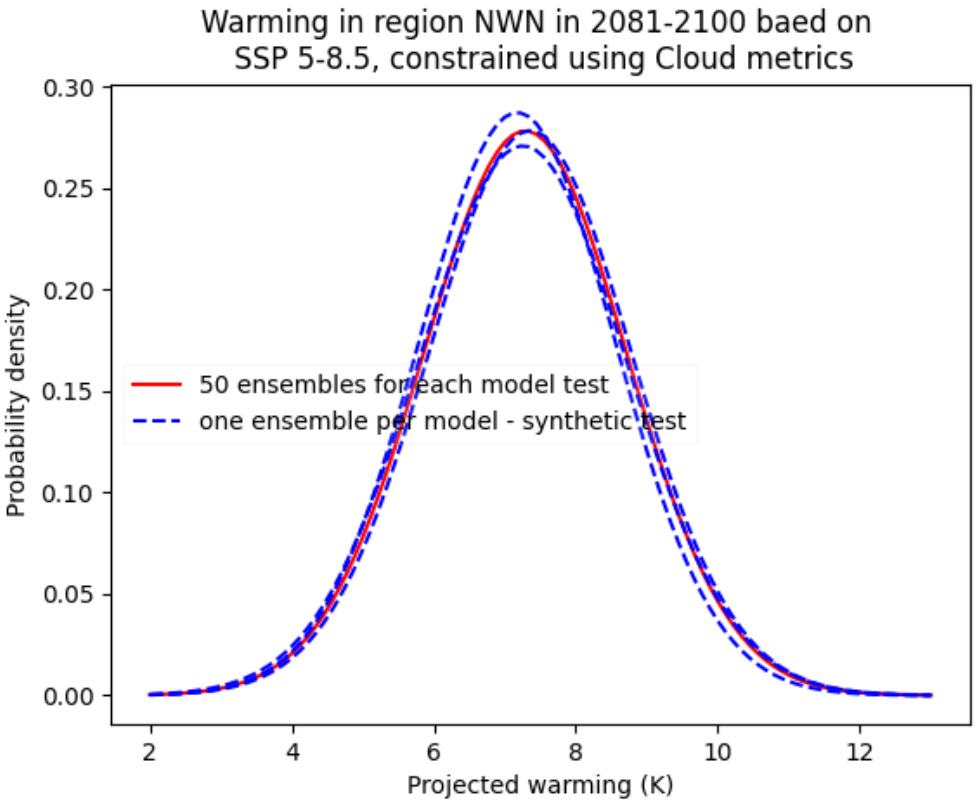


Figure AC.S1 Constrained warming in synthetic data experiment (described in Text S1). The PDFs are constrained projections for global mean surface air temperature change (2081-2100 relative to 1995-2014 under SSP 5-8.5) by applying MBLC and BCS over NWN (N.W.North-America). The dashed red PDF is derived using the approach used in the main manuscript, but sampling from synthetic 50-member ensembles from each model (see Step 2 of Text S1). The blue PDFs are derived from the same data, but first restricting ensemble sizes for each model to the number of simulations actually available (see Step 3 of Text S1).

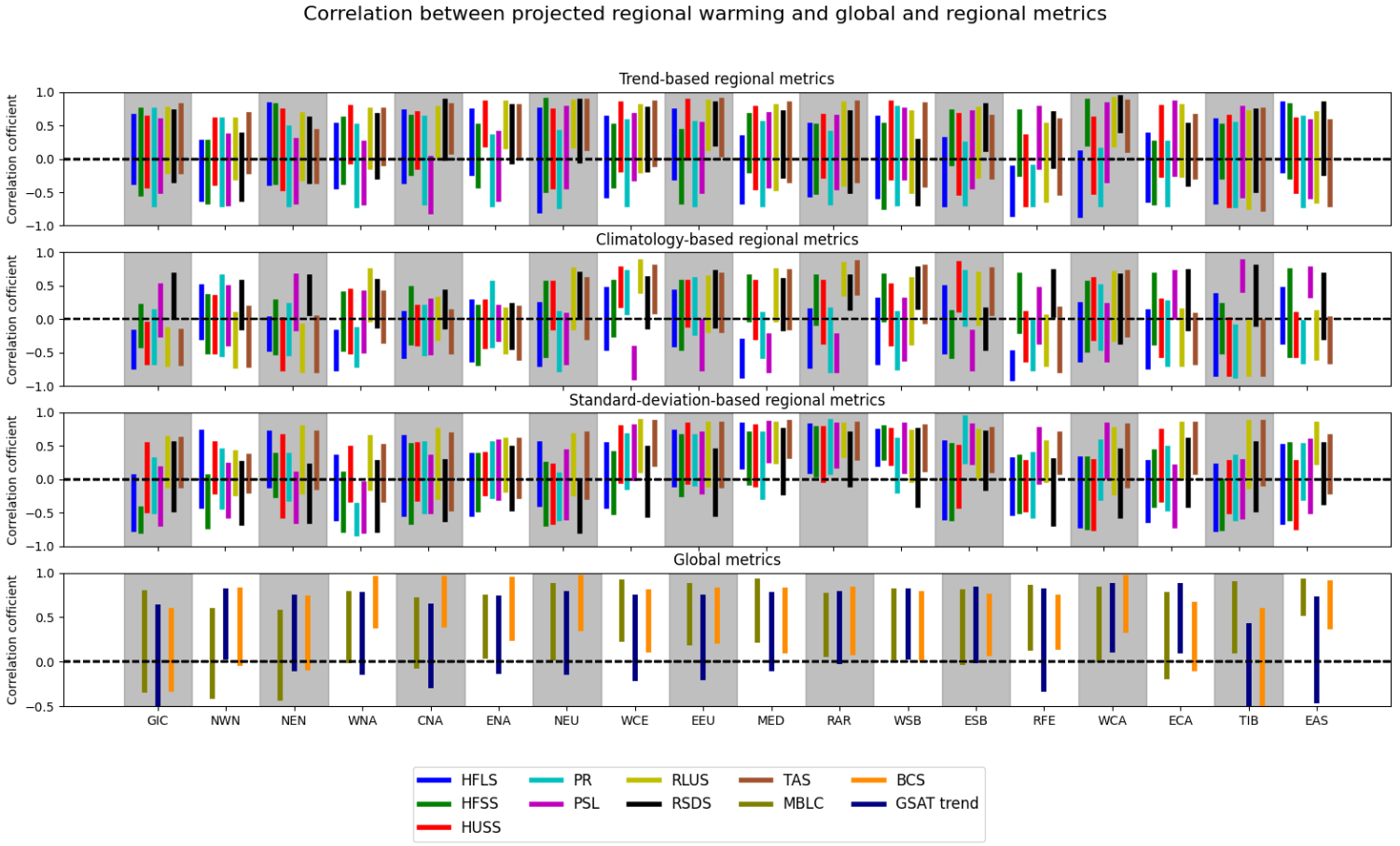


Figure AC.S2 Same as Fig 4.2, but for CMIP5 ensembles.

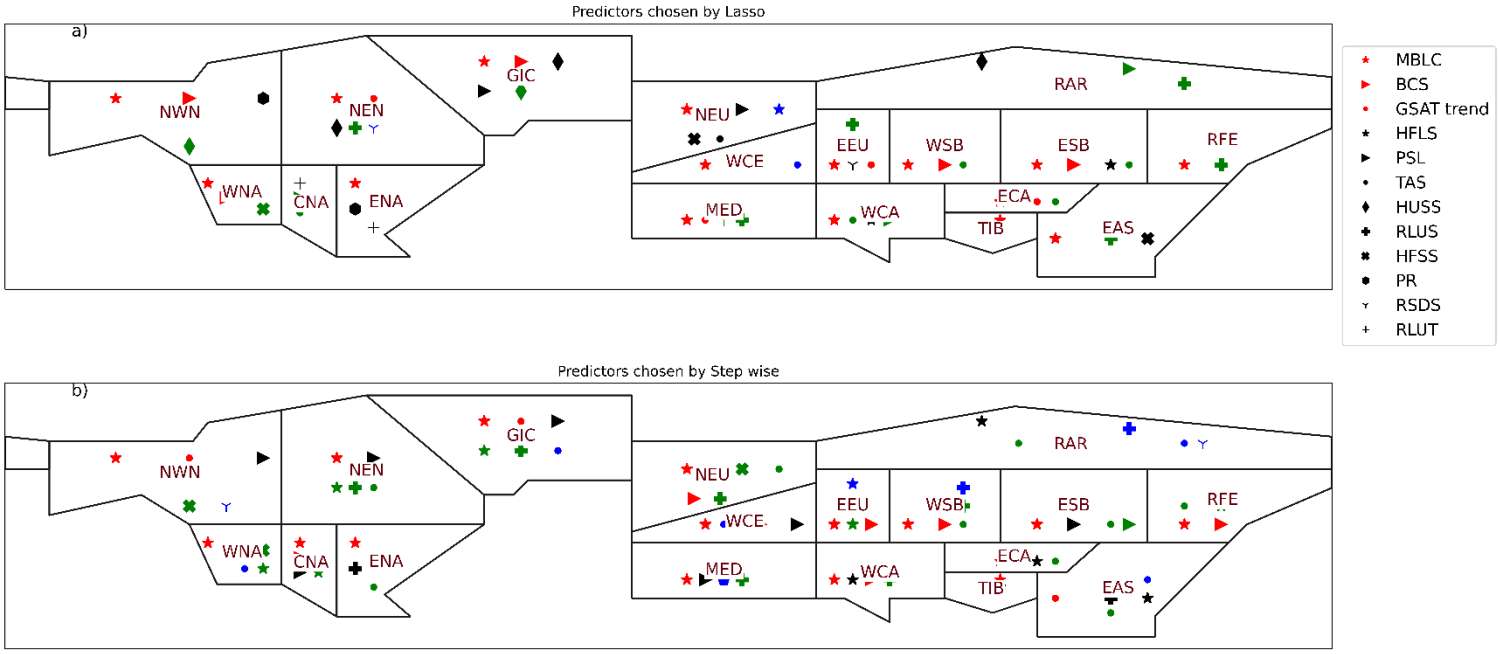


Figure AC.S3 The predictors selected for each region based on the metric selection process. Panel (a) indicates predictors selected by the Lasso selection strategy and Panel

(b) indicates predictors selected by the stepwise selection strategy. Predictors are described in Table 4.2. Markers in red indicate global metrics while the markers in black, blue and green indicate trend-based, climatology-based and standard-deviation-based regional metrics respectively.

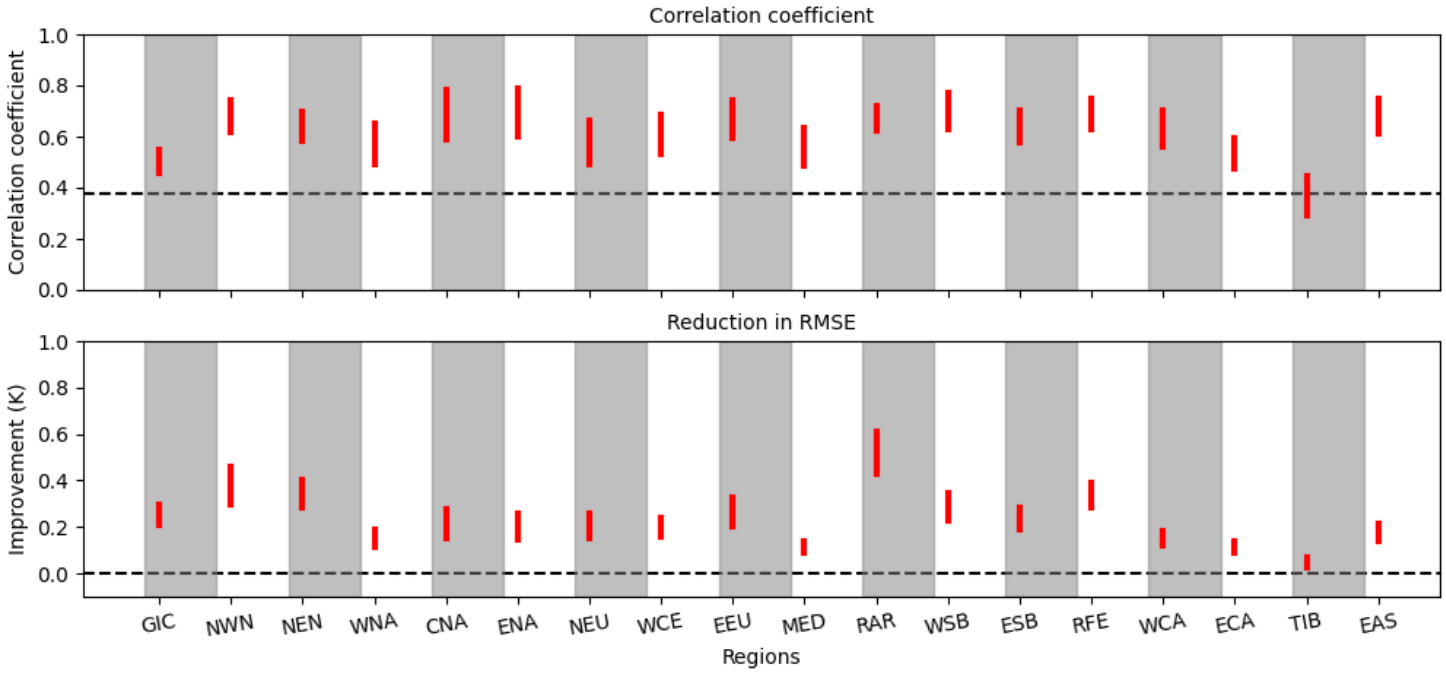


Figure AC.S4 As in Fig 4.3, but for CMIP6 SSP 1-2.6 using cloud metrics.

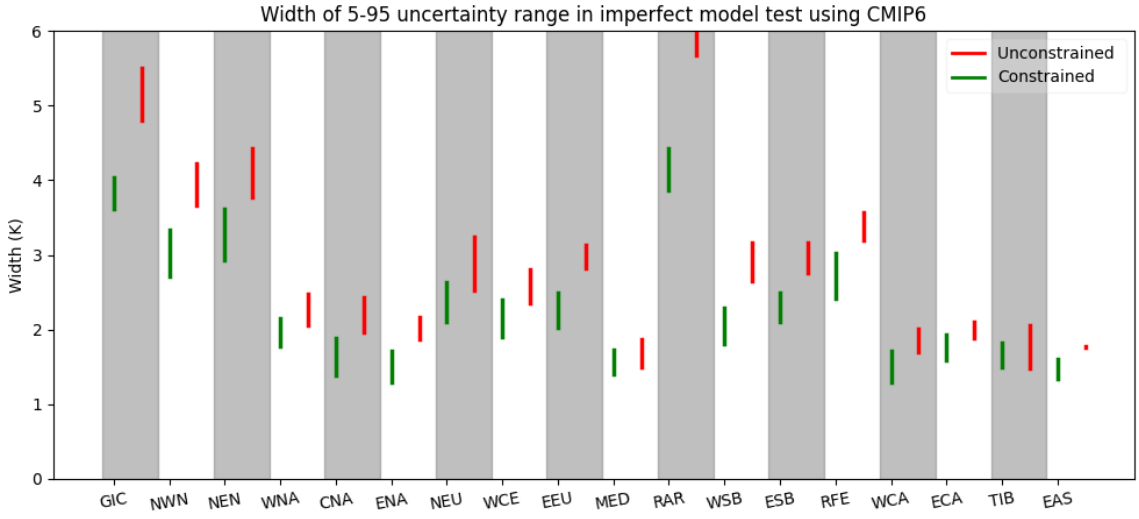


Figure AC.S5 As in Fig 4.4, but for CMIP6 SSP 1-2.6 using cloud metrics.

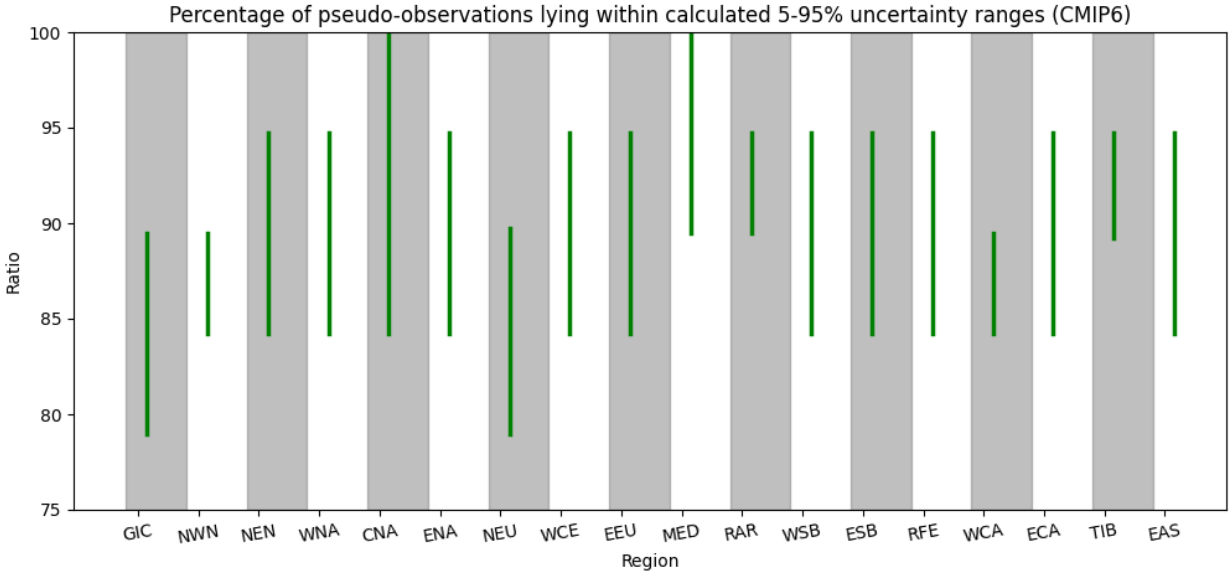


Figure AC.S6 As in Fig 4.5, but for CMIP6 SSP 1-2.6 using cloud metrics.

Appendix D

Text S1. As described in Section 5.2.2, our study samples over initial condition ensembles. Each sample gives us a distribution of projected GSAT change [noted as y in below Eq. (1)] with a particular mean (μ) and standard deviation (σ). Hence, multiplying the joint distribution of these two statistics $f(\mu, \sigma)$, by a conditional distribution $f(y|\mu, \sigma)$, and then integrating over μ and σ, gives us a population estimate $p(y)$ of the marginal PDF of projected GSAT changes:

$$p(y) = \iint f(y|\mu,\sigma) f(\mu,\sigma)\, d\mu\, d\sigma \qquad (1)$$

Based on our sampling strategy, each derived PDF with its corresponding value of μ and σ is equally probable. Sampling μ and σ from their joint distribution and then averaging the resulting conditional distributions of y gives a sample estimate of this population mean distribution.

Text S2. In order to determine the optimal time period to minimize the influence of internal variability and maximize the strength of the statistical relationship between past trends and future changes in the CMIP6 models, we vary the date range over which trends are calculated. We calculate the correlation (r) between historical trends (based on different initial years and the final years) and projected warming between 1995-2014 and 2081-2100 under SSP5-8.5. In this calculation, we consider the first ensemble member of each model. The corresponding results are shown in Fig AD.S6.

As in Fig S6 (a), the early periods (e.g when we hold the end year at 1999 and vary the start year) do not indicate a strong emergent relationship, probably due to large inter-model differences in the response to aerosol emissions changes over these periods which are expected to confound such a relationship (Chapter 2;Jimenez-de-la-Cuesta; Mauritsen 2019; Nijsse et al. 2020). Holding the end year close to the present (e.g. 2014 or 2022) and making the start year after the in the 1970s or later results in strong correlation coefficient except for the periods 1999-2014/2022 and 2014-2022, which are too short, with a strong influence of internal variability. Likely since the GSAT response to uncertain aerosol forcing is relatively constant over this period (Chapter 2; Nijsse et al. 2020), the GSAT trend from the 1970s to the present is better correlated with future projected warming compared with the centennial trend (e.g. 1850-2022). As shown in Fig S6 (b) and (c), by reducing the internal variability, removing the ETP-congruent part of the GSAT trend results in a stronger correlation coefficient, especially for GSAT trends calculated over a relatively short period (e.g. 1999-2014, 1999-2022 and 2014-2022).

The periods for GSAT trend with ETP variability removed that give optimal correlation coefficient are 1970-2022 (with correlation coefficient 0.82), 1984-2014 (0.83), 1984-2022 (0.81), and 1999-2022 (0.81). These periods all show strong correlation between past trends and future changes across models. Considering the long period is expected to reduce the influence of internal variability (and we note that results shown in Figure

S6 are based on a single ensemble member), we therefore take 1970-2022 as the optimal period to constrain future warming (section 5.3.4). For comparison, we also compare with the results over a relatively short period, 1993-2012, over which the variability of ETP trend is expected to play an event stronger role.

Table AD.S1 List of CMIP6 Historical, SSP5-8.5 and SSP1-2.6 simulations used in this paper. The numbers of ensemble members used for each experiment are listed in the second, the third and the fourth columns. We use all simulations for which the necessary model output was available. For the historical trend based on 1970-2022, we use the same set of simulations throughout, and discard historical simulations for which a corresponding SSP simulations is not available

Model name	Historical	SSP5-8.5	SSP1-2.6
ACCESS-CM2	1	1	1
ACCESS-ESM1	10	10	10
BCC-CSM2-MR	3	1	1
CanESM5	50	50	50
CAMS-CSM1-0	1	1	1
CAS-ESM2-0	2	2	2
CESM2	6	2	2
CNRM-CM6-1	10	6	6
CNRM-ESM2-1	5	5	5
FGOALS-f3-L	3	1	1
FGOALS-g3	1	1	1
GFDL-ESM4	1	1	1
GISS-E2-1-G	13	13	13
GISS-E2-1-H	10	10	10
HadGEM3-GC31-LL	4	4	4
HadGEM3-GC31-MM	4	4	4
INM-CM5-0	1	1	1
IPSL-CM6A-LR	32	7	7
KACE-1-0-G	3	2	2
MIROC-ES2L	7	7	7
MIROC6	50	50	50
MPI-ESM1-2-HR	5	1	1
MPI-ESM1-2-LR	8	8	8
MRI-ESM2-0	6	6	6
NESM3	4	4	4
NorESM2-LM	3	1	1
NorESM2-MM	3	1	1
UKESM1-0-LL	8	4	4

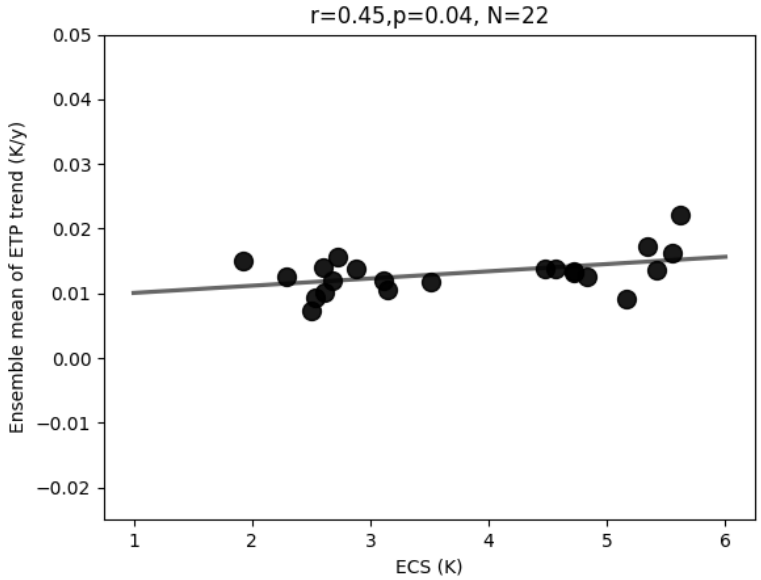


Figure AD.S1 Scatter plot showing ensemble mean of 1970-2022 ETP trend versus ECS for each model, and the corresponding regression fit. 'r' represents the correlation coefficient between ECS and the ETP trend across the multi-model ensemble with the corresponding 'p' value, while 'N' represent the number of climate models in use.

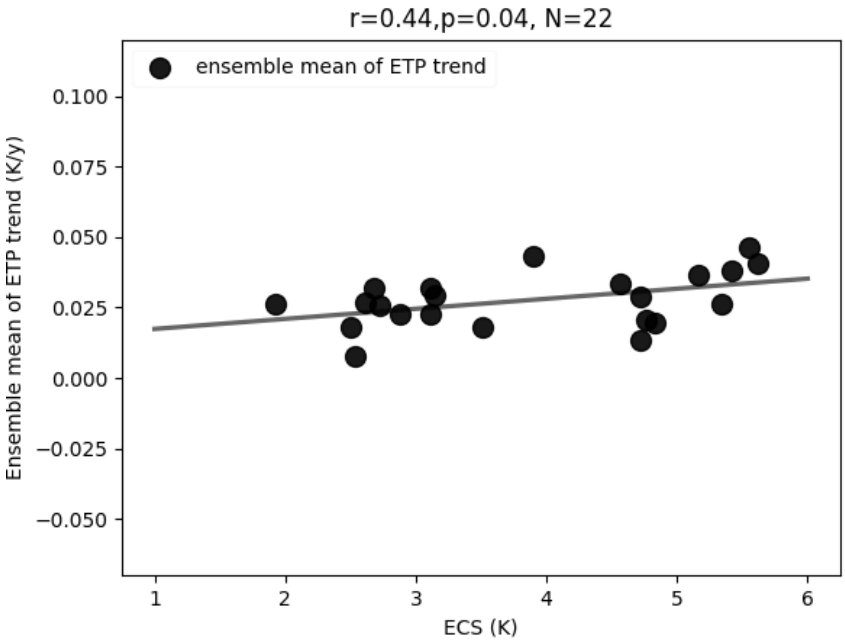


Figure AD.S2 Similar to Fig S1, but for the period of 1993-2012.

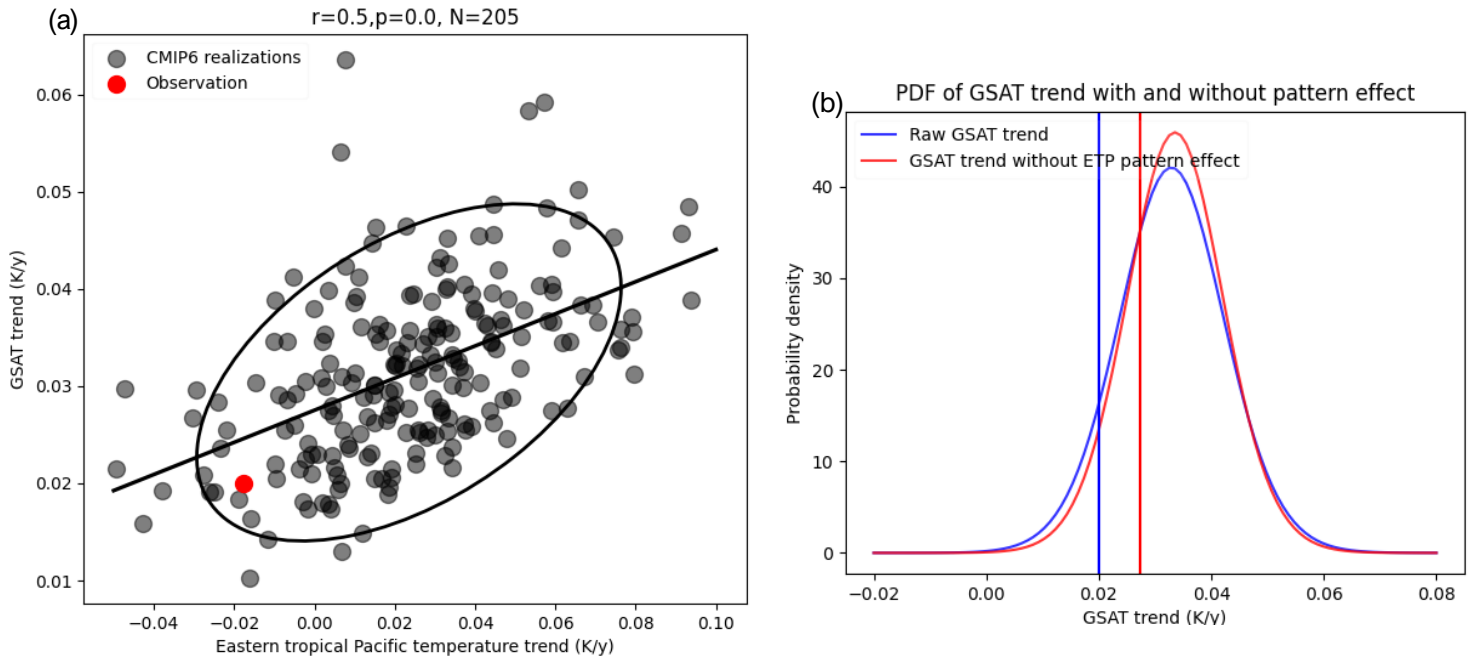


Figure AD.S3 As Fig 5.2 but for the period 1993-2012.

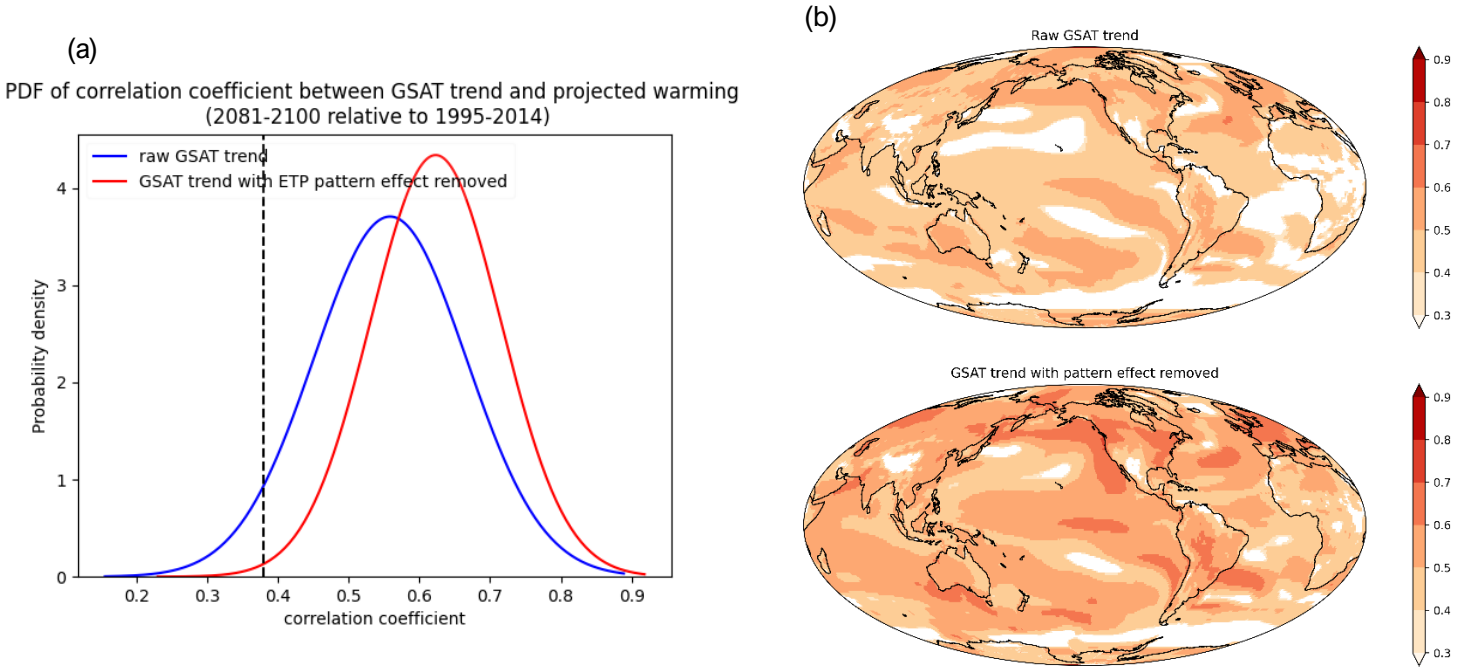


Figure AD.S4 As Fig 5.3 but for the period of 1993-2012.

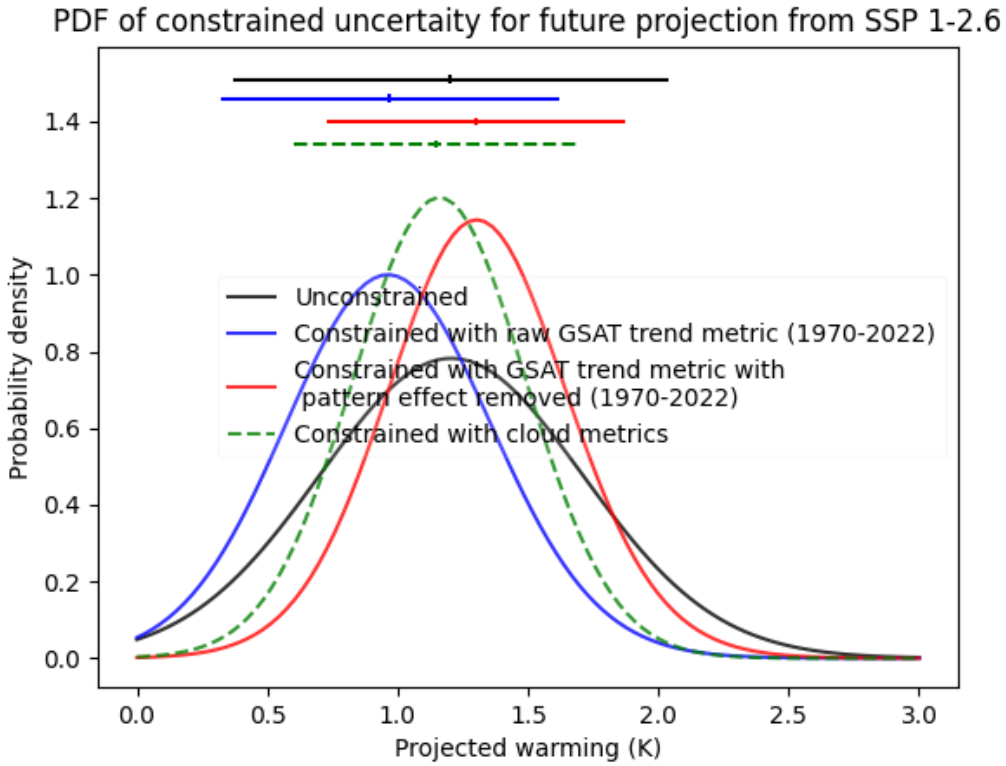


Figure AD.S5 As Fig 5.5 but based on SSP 1-2.6.

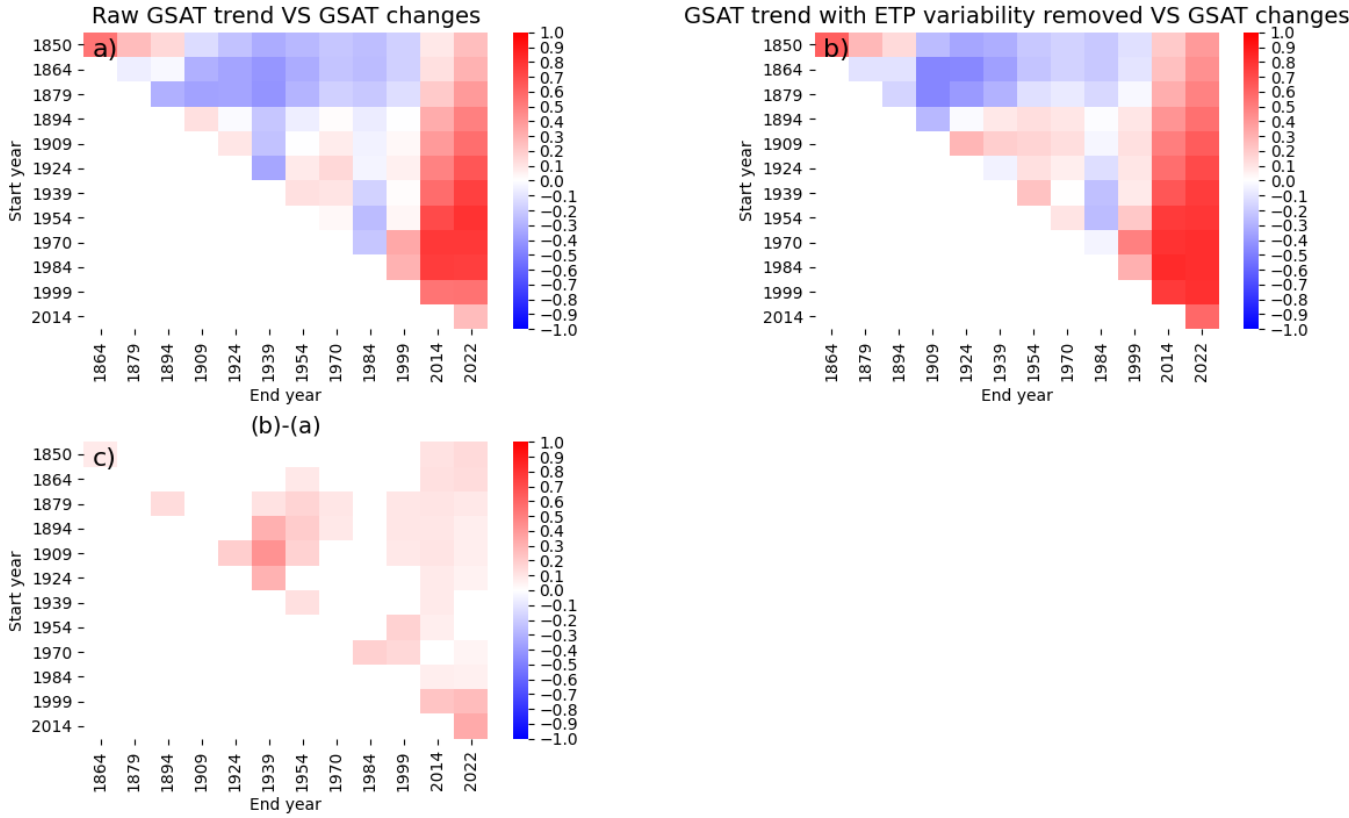


Figure AD.S6 The correlation coefficient between the historical GSAT trend (based on different initial years and the different final years) and future projected warming (between 1995-2014 and 2081-2100 under SSP5-8.5), based on one ensmeble per

model. The *y-axis* represents the start years while the *x-axis* represents the end years. Panel (a) shows the correlation coefficient between raw GSAT trend and projected warming while panel (b) shows the correlation coefficient between GSAT trend with ETP variability removed. Panel (c) shows the difference between panel (b) and (a). The shading areas in panel (c) represent two correlation coefficients that are significantly different from each other at the p=0.05 level.