To what extent does discounting ‘hot’ climate models improve the predictive skill of climate model ensembles?

Abigail McDonnell*[1], Adam Michael Bauer[2], and Cristian Proistosescu[1,3]

[1] *Department of Climate, Meteorology, and Atmospheric Sciences, University of Illinois Urbana-Champaign, 1301 W Green St, Urbana IL 61801*

[2] *Department of Physics, University of Illinois Urbana-Champaign, 1110 W Green St, Loomis Laboratory, Urbana, IL 61801*

[3] *Department of Earth Sciences and Environmental Change, University of Illinois Urbana-Champaign, Urbana, IL 61801*

Forthcoming in *Earth’s Future*

September 20, 2024

Abstract

It depends. The Intergovernmental Panel on Climate Change’s (IPCC) Assessment Report Six (AR6) took a step towards ending so-called ‘model democracy’ by discounting climate models that are too warm over the historical period (i.e., models that ‘run hot’) when making projections of global temperature change. However, the IPCC did not address whether this procedure is reliable for other quantities. Here, we explore the implications of weighting climate models according to their skill in reproducing historical global-mean surface temperature using three other climate variables of interest: annual average precipitation change, regional average temperature change, and regional average precipitation change. We find that the temperature-based weighting scheme leads to an improved prediction of global average precipitation, though we show that this prediction could be overconfident. On regional scales, we find a heterogeneous pattern of error reduction in future regional precipitation. This stands in sharp contrast with

*Corresponding author email: amm18@illinois.edu

the broad regional pattern of error reduction in future temperature projections, though we do find regions where error is not significantly reduced. Our results demonstrate that practitioners using weighted climate model ensembles for climate projections must take care when weighting by temperature alone, lest they produce unreliable climate projections that result from an inappropriate weighting procedure.

Keywords: Climate change, climate projections, CMIP6, model democracy

Plain Language Summary

Climate model ensembles are widely used for risk assessment. However, a few of the most recent generation climate models 'run hot' in the historical period, widening the spread of future global warming. The Intergovernmental Panel on Climate Change's (IPCC) sixth assessment report presents a number of weighting schemes to address this 'hot model' problem, each of which discount models that are 'too hot' in the historical period. However, it is unclear if this procedure is reliable for other quantities of interest. Here we explore the impact of this procedure on global average precipitation change, regional temperature change, and regional precipitation change. We find that while this scheme improves the prediction of global precipitation change and generally improves the prediction of regional temperature, it does not broadly improve regional predictions of future precipitation change. We conclude that users of climate model output must be careful when applying a global temperature-based weighting scheme in regional impact studies.

Key points

- Using historical warming to weight climate models can improve global predictions of annual temperature change and precipitation change.
- Using past warming to weight future climate projections has varied effects on regional error reduction depending on the metric of interest.
- Climate model end-users should use caution when applying a weighting scheme to avoid biased or overconfident assessments of climate impacts.

1 Introduction

The Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al., 2016) includes numerous updates to physical processes that have substantially broadened the range of equilibrium climate sensitivity (ECS) to include values much higher than previous CMIP generations (Zelinka et al., 2020; Forster et al., 2020). These more sensitive models also simulate more end-of-century warming and have been criticized for 'running hot' (Hausfather et al., 2022). 'Hot models' in CMIP6 generally have two common biases: (i) they simulate too much warming over the last four decades because their transient climate response (TCR) is outside generally accepted values (Hausfather et al., 2022), and (ii) they have an unrealistically large estimate of ECS relative to state-of-the-art estimates (Sherwood et al., 2020).

In response to this bias, the Intergovernmental Panel on Climate Change's (IPCC) Sixth Assessment Report (AR6) started down-weighting 'hot models' when providing projections of global-mean temperature changes (Eyring et al., 2021). This model weighting method ended so-called 'model democracy' present in previous CMIP generations (Knutti, 2010; Brunner et al., 2019), in which all models are given equal weight in computing the ensemble average of a given climate variable (i.e., end-of-century temperature rise). To summarize the IPCC's approach, they used a percentile-by-percentile average of three distinct weighting schemes (Tokarska, 2020; Liang et al., 2020; Ribes et al., 2021), each of which discount models that are 'too hot', to form a constrained future projection of relative global surface air temperature change, and then utilized an emulator to generate future projections (see Figure 4.11 and supplemental data in Eyring et al., 2021). One method present in AR6, Liang et al. (2020), show that their weighting approach reduces overall bias in future global-mean temperature projections (via cross validation), and therefore provides a more precise estimation of *global-mean* warming.

It is unclear, however, if weighting climate model projections by historical global-mean warming trends has skill for quantities other than global-mean warming, specifically those that may not be well correlated with global temperature changes (Hausfather et al., 2022). For example, while global temperature changes have been shown to correlate with *global* precipitation changes, the correlation appears to be strong only in the polar regions (Shiogama et al., 2022), suggesting that weighted ensembles might not offer a more skillful prediction for precipitation in the mid- and low-latitudes.

AR6 itself offers no guidance for weighting quantities other than global temperature. The proposed workaround – using global warming levels – is unfortunately not informative for any estimates of impacts or risk that require time-horizons, broadly classified as 'transition risks' (Bauer et al., 2024b). For example, adjustment costs (Lucas, 1967; Mussa, 1977) and economic inertia (Ha-Duong et al., 1997) link the cost of abating CO_2 emissions to the rate of abatement and have been shown to imply much more aggressive near-term climate policies (Campiglio et al., 2022; Bauer et al., 2024a). Therefore, the rate of warming – which directly influences the optimal rate of abatement in integrated assessment models with adjustment costs – is an important consideration for policymakers. Worse still, it is unclear if a better prediction globally implies a uniformly better prediction on regional scales, or if the bulk of the precision is gained in locations relatively uninteresting for a specific impact analysis (i.e., in polar regions, as opposed to the low- and mid-latitudes).

Here we demonstrate the issues of weighting models according to their skill in reproducing global-mean surface temperature using three other climate variables of interest: annual average precipitation, regional temperature change, and regional precipitation change. We use a weighting scheme that is most similar to Liang et al. (2020), which itself expands on the weighting scheme outlined in Knutti et al. (2017). The general approach is to compute model weights using a model's ability to replicate historical warming, while also accounting for model interdependency (see the *Supplementary Materials* for more details). The ability for a given weighting scheme to reduce out-of-sample prediction error is evaluated via a perfect model test (*Supplemental Materials*), which can be summarized as: (i) a model is randomly chosen as truth, and referred to as the 'pseudo-observation'; (ii) the other models are weighted based on their ability to reproduce the historical period in the pseudo-observation; (iii) the weighted ensemble projections are compared with 21$^{\text{st}}$ century predictions from the model chosen as pseudo-observations. This procedure is carried out with each ensemble member as the pseudo-observation once to produce the distribution of 'perfect model test errors' seen in Figure 1 and to compute the change in RMSE in Figure 2.

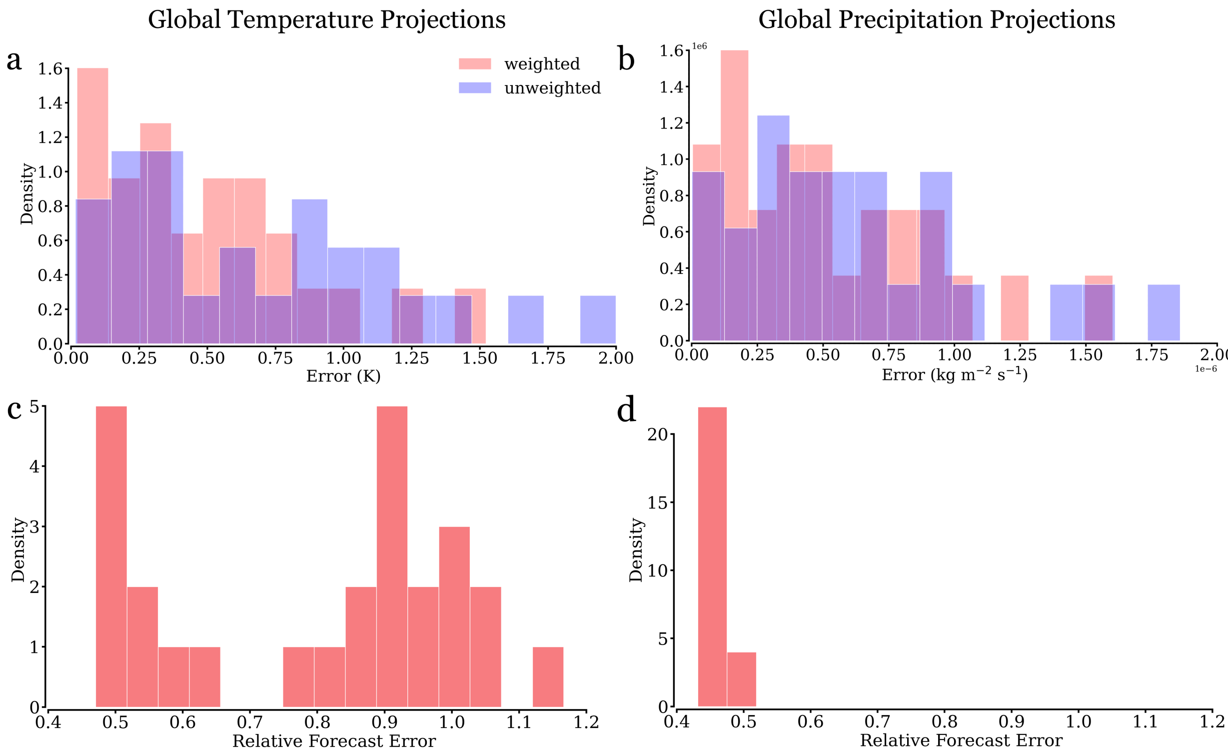

Figure 1: Perfect model test error and relative forecast error distributions. Panel **a** shows a histogram of the perfect model test errors in projections of 21$^{\text{st}}$ global warming using weighted (red) and unweighted (blue) distributions. Panel **b** is as Panel **a**, but for the error in global precipitation projections using temperature-based weights. Panel **c** shows the histogram of relative forecast error for global temperature projections using temperature-based weights. Panel **d** is as Panel **c** but for global precipitation projections using temperature-based weights.

2 Results

2.1 Global Analysis

We first apply the temperature-based weighting scheme to projections of 21st century global average temperature change, which is analogous to AR6's analysis. We show the distribution of perfect model test errors in Figure 1**a**, which reflects the distribution of error for all pseudo-observation choices. We find that applying the temperature-based weights reduces root mean squared error (RMSE) between the weighted and unweighted distribution by 25.4% and reduces the relative forecast error (RFE; the ratio of the weighted ensemble variance and the unweighted ensemble variance, see the *Supplementary Materials*) by 18%. This reduction in RMSE suggests that weighting models by their ability to reproduce historical warming results in a more reliable prediction of future global-mean temperature. Likewise, a reduction in RFE implies that the spread of the weighted future projection is less than the unweighted ensemble, which naturally follows from the weighting scheme discounting models that are dissimilar to the global temperature trend of the chosen pseudo-observation. Our findings are consistent with those found in Liang et al. (2020), and corroborate the idea that historical temperature trends can be used to constrain global temperature projections.

Next, we apply the historical temperature-based weighting scheme to global precipitation projections. We find that weighting future precipitation projections by historical warming trends reduces RMSE by 17.8% and decreases RFE by 55%. This suggests that temperature-based weights are useful in constraining global-mean precipitation projections, but could introduce overconfidence; indeed, the RFE is reduced three times more in the precipitation projections than in the temperature projections. The temperature-based weighting scheme likely decreases global precipitation RMSE and RFE because global temperature trends and global precipitation trends have been shown to be well-correlated (Shiogama et al., 2022). This correlation may be explained by the fact that models may agree on global precipitation changes (Held and Soden, 2006), while nonetheless disagreeing on the specific pattern of precipitation and precipitation change.

Judging by the results for the global-mean, one may be tempted to conclude that historical temperature trends are a reasonable predictor of future precipitation anomalies in CMIP6. This result is consistent with past work framing historical temperature as a possible emergent constraint

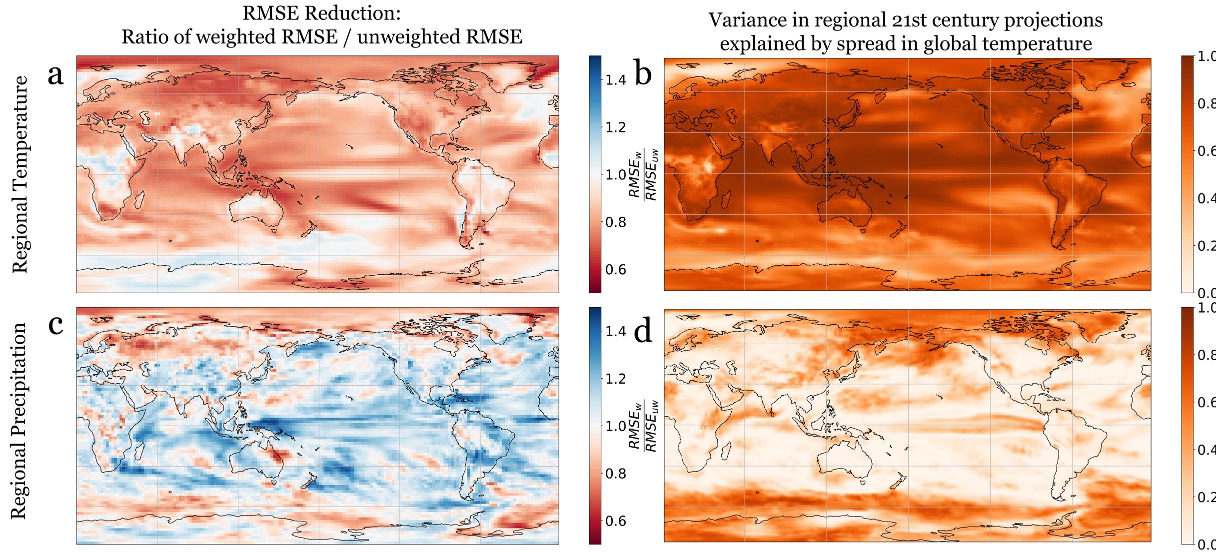

Figure 2: Regional decomposition of RMSE reduction and variance explained by weighting metric. Panel **a** shows the spatial distribution of the relative RMSE between the raw ensemble mean and the historical temperature trend-derived weighting technique, averaged over each pseudo-observation choice. Here, a value less than (greater than, resp.) unity implies a more (less, resp.) precise prediction using the weighted ensemble mean as opposed to the unweighted ensemble mean. Panel **b** shows the variance in future regional temperature anomalies that is explained by historical global temperature trends. Panel **c** is as Panel **a**, but for precipitation projections weighted by historical temperature trends. Panel **d** shows the variance in future regional precipitation anomalies explained by historical global temperature trends. Note that high levels of variance explained should correspond to a relative RMSE of less than one.

on future global average precipitation projections (Shiogama et al., 2022). But does it follow that climate model practitioners will find this procedure useful in their impact analysis for, say, an individual city? To answer this question, we next look at regional decompositions of RMSE changes.

2.2 Regional Analysis

We find that weighting models by global-mean historical warming trends produces a well-defined pattern of RMSE reduction for regional warming projections (Figure 2**a**). This pattern can be explained by the robust pattern of correlation between the ensemble spread in *global-mean* 21$^{\text{st}}$ century temperature trends and the ensemble spread in *regional* 21$^{\text{st}}$ century temperature trends (Figure 2**b**). The high degree of correlation over most of the world implies that future regional temperature trends are robustly predicted by future global-mean temperature trends, which are themselves constrained by historical global-mean temperature trends.

The locations where regional projections are least improved by weighting correspond to locations where the correlation between global and regional trends is low. These are locations where surface temperature is strongly controlled by local ocean processes, such as parts of the Southern Ocean and the North Atlantic. The implication is that in these regions the uncertainty in local warming trends is not primarily determined by uncertainty in global warming, but rather by regional ocean dynamics.

In stark contrast, we find that weighting regional precipitation trends by how well models reproduce historical global warming does not lead to wide-spread reduction in RMSE (Figure 2**c**). This finding can largely be attributed to the lack of correlation between *global-mean temperature trends* and *regional-mean precipitation trends* over the 21$^{\text{st}}$ century (Figure 2**d**) outside of the polar regions.

This low degree of correlation suggests that outside the polar regions, uncertainty in regional precipitation is not dominated by the same processes that determine uncertainty in global warming trends. Indeed, the future regional precipitation in polar regions is likely well-correlated with global temperature because it is primarily 'thermodynamically controlled', whereas changes in regional precipitation in the mid-latitudes and tropics are 'dynamically controlled' and as such the non-polar regional precipitation trends cannot be easily linked to global temperature rise (Emori and Brown, 2005). This could also explain why regional RMSE change in future precipitation projections is heterogeneous, while global projections are improved: while models all agree that precipitation increases with temperature (Held and Soden, 2006), they might disagree on the (dynamically-controlled) location of precipitation changes. Moreover, owing to dynamical differences of how individual models represent precipitation, regional precipitation patterns may not be robust prior to weighting, meaning the changes after weighting are also not robust.

In any event, the lack of predictive power of historical global temperature anomalies for regional precipitation projections make this metric a poor choice to weight climate models on regional scales (at least, for precipitation projections). As a quantitative example, average RMSE over the contiguous United States (CONUS) in future temperature is reduced by 20%, while average RMSE in future precipitation over the CONUS is increased by 8%. (We provide the same calculations for 17 additional regions in the *Supplementary Materials* and find similar results.)

3 Discussion

Our regional analysis shows that care must be taken when applying a historical temperature-based weighting scheme to other climate variables, particularly for regional impact analysis. Weighting regional temperature projections by global historical temperature projections leads to a reduction in bias for most regions. However, the same is absolutely not true when applying these weights to regional precipitation projections. Indeed, we find that many regions actually have a higher RMSE in future precipitation projections using the weighted mean as opposed to the unweighted mean (see blue regions in Figure 2**c**). It also follows that a global reduction in bias cannot be conflated with a useful, uniform regional reduction in error; indeed, just because error decreases globally does not imply it also decreases in every city or municipality, or even in most cities or municipalities. We conclude that, based on these results, the skill of a temperature-based weighting scheme for global temperature projections cannot be generalized across different climate metrics.

These results have important implications for both climate model practitioners and climate scientists. For practitioners, it is important to take caution when choosing a weighting scheme so as to avoid an overconfident or biased prediction. Naïvely applying a temperature-based weighting method, such as those adopted by the IPCC, to other climate model variables can lead to misleading results and worse regional predictions than using an unweighted mean, as shown in Figure 2. (Note we also explore a historical precipitation-based analog to the IPCC's temperature-based weighting scheme in the *Supplementary Materials* and find similar conclusions.) In other words, temperature-based weighting methods cannot, and should not, be considered general for any regional metrics other than temperature projections without verification on a case-by-case basis. Note that applying an unweighted mean in many cases may still not be preferable, and one would ideally use a weighting scheme that is optimized for the region and variable of interest. Our work would suggest that a reliable first-step would be to determine the degree to which a candidate global weighting metric correlates to the regional climate variable of interest (i.e., our calculations in Figure 2**b,d**) to probe if a weighted projection would improve bias in a climate model ensemble. Note that our approach is generalizable to the case where multiple candidate weighting metrics correlate well with a desired climate variable (see the *Supplementary Materials*).

More work needs to be done to build out robust, bespoke weighting schemes for different metrics of climate change, particularly those that are important for climate impact analysis. Weighting schemes for multi-faceted climate risk assessment, where multiple variables need to be predicted, are likely to be particularly challenging. Different impact variables could correlate well with different weighting metrics, but using different weighting schemes for each impact variable could lead to inconsistent predictions. If it is possible to define a single weighting metric that correlates well with each climate variable of interest, we recommend that weighting schemes are built around this weighting metric. Additional approaches to address this issue have been suggested, such as recent work focusing on choosing climate models based on independence, performance, and spread (Merrifield et al., 2023); expanding on this framework or developing novel approaches would fill a need for model weighting and selection for specific tasks, particularly on regional spatial scales and near-term timescales, where reliable risk assessment is urgently needed (Condon, 2023). Climate scientists should actively engage with climate model practitioners to lend expertise and insight into the best practices for weighting climate ensembles, lest highly consequential decisions be made based on ill-suited (though well-intentioned) weighted climate model ensembles.

Disclosure Statement

AM: I have no conflicts of interest to disclose.

AMB: I have no conflicts of interest to disclose.

CP: I have no conflicts of interest to disclose.

Open Research

The raw climate model data listed in the *Supplementary Materials* is publicly available in a long-term stable repository at https://esgf-data.dkrz.de/search/cmip6-dkrz/. We have compiled a text list of models to download from the above repository. The code and stable repository model download list to reproduce all analysis can be found in McDonnell (2024) and at the following Github link: https://github.com/abigailmcdonnell/model-democracy.

Author ORCIDs

AM: 0000-0001-9749-9474

AMB: 0000-0002-7471-8934

CP: 0000-0002-1717-124X

Acknowledgements

The authors would like to thank Ryan Sriver, three anonymous reviewers, and the attendees of the fall meeting of the American Geophysical Union 2022 for providing useful feedback and discussions. AM and CP acknowledges support from a National Oceanic and Atmospheric Sciences grant No. OAR MAPP UWSC12184. AM acknowledges support from the National Oceanic and Atmospheric Sciences Hollings Scholarship Program. AMB acknowledges support from a National Science Foundation Graduate Research Fellowship grant No. DGE 21-46756 and the Climate Support Facility of the World Bank Group. Computations were performed on the Keeling computing cluster, a computing resource operated by the School of Earth, Society and Environment (SESE) at the University of Illinois Urbana-Champaign. We would also like to thank Ryan Abernathey's writings for providing a template for using CMIP6 data via a Google Cloud repository and the Pangeo Stack (Abernathey, 2024).

Author Contributions

CP conceived the study. AM wrote the code, gathered and analyzed data, made the figures, and wrote the first draft of the paper. AMB developed the computational approach and technical details. AMB and CP provided guidance and advising. All authors assisted in editing this draft and approve the submitted version.

References

R. Abernathey. CMIP6 in the Cloud Five Ways, Dec. 2024. URL https://medium.com/pangeo/cmip6-in-the-cloud-five-ways-96b177abe396.

A. M. Bauer, F. McIsaac, and S. Hallegatte. How Delayed Learning about Climate Uncertainty Impacts Decarbonization Investment Strategies. Working Paper WPS10743, The World Bank Group, Washington DC, Mar. 2024a. URL https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099829103282438373/idu1f2d86d77127091490d1a6df1dc342f15d10b.

A. M. Bauer, C. Proistosescu, and G. Wagner. Carbon Dioxide as a Risky Asset. *Climatic Change*, 177(5):72, May 2024b. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-024-03724-3. URL https://link.springer.com/10.1007/s10584-024-03724-3.

L. Brunner, R. Lorenz, M. Zumwald, and R. Knutti. Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, 14(12):124010, Dec. 2019. ISSN 1748-9326. doi: 10.1088/1748-9326/ab492f. URL https://iopscience.iop.org/article/10.1088/1748-9326/ab492f.

E. Campiglio, S. Dietz, and F. Venmans. Optimal Climate Policy as If the Transition Matters. Working Paper 10139, CESifo, Munich, Germany, 2022. URL https://www.cesifo.org/en/publications/2022/working-paper/optimal-climate-policy-if-transition-matters.

M. Condon. Climate Services: The Business of Physical Risk. *Arizona State Law Journal*, 55(1): 147, Apr. 2023. URL https://scholarship.law.bu.edu/faculty_scholarship/3658.

S. Emori and S. J. Brown. Dynamic and thermodynamic changes in mean and extreme precipitation under changed climate. *Geophysical Research Letters*, 32(17):2005GL023272, Sept. 2005. ISSN 0094-8276, 1944-8007. doi: 10.1029/2005GL023272. URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2005GL023272.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, May 2016. ISSN 1991-9603. doi: 10.5194/gmd-9-1937-2016. URL https://gmd.copernicus.org/articles/9/1937/2016/.

V. Eyring, N. Gillett, K. Achuta Rao, R. Barimalala, M. Barreiro Parrillo, N. Bellouin, C. Cassou, P. Durack, Y. Kosaka, S. McGregor, S. Min, O. Morgenstern, and Y. Sun. *Human Influence on the Climate System*, page 423–552. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896.005.

P. M. Forster, A. C. Maycock, C. M. McKenna, and C. J. Smith. Latest climate models confirm need for urgent mitigation. *Nature Climate Change*, 10(1):7–10, Jan. 2020. ISSN 1758-

678X, 1758-6798. doi: 10.1038/s41558-019-0660-0. URL https://www.nature.com/articles/s41558-019-0660-0.

M. Ha-Duong, M. J. Grubb, and J.-C. Hourcade. Influence of socioeconomic inertia and uncertainty on optimal CO_2-emission abatement. *Nature*, 390(6657):270–273, Nov. 1997. ISSN 0028-0836, 1476-4687. doi: 10.1038/36825. URL http://www.nature.com/articles/36825.

Z. Hausfather, K. Marvel, G. A. Schmidt, J. W. Nielsen-Gammon, and M. Zelinka. Climate simulations: recognize the 'hot model' problem. *Nature*, 605(7908):26–29, May 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/d41586-022-01192-2. URL https://www.nature.com/articles/d41586-022-01192-2.

I. M. Held and B. J. Soden. Robust Responses of the Hydrological Cycle to Global Warming. *Journal of Climate*, 19(21):5686–5699, Nov. 2006. ISSN 1520-0442, 0894-8755. doi: 10.1175/JCLI3990.1. URL http://journals.ametsoc.org/doi/10.1175/JCLI3990.1.

R. Knutti. The end of model democracy?: An editorial comment. *Climatic Change*, 102(3-4): 395–404, Oct. 2010. ISSN 0165-0009, 1573-1480. doi: 10.1007/s10584-010-9800-2. URL http://link.springer.com/10.1007/s10584-010-9800-2.

R. Knutti, J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring. A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, 2017. ISSN 00948276. doi: 10.1002/2016GL072012. URL http://doi.wiley.com/10.1002/2016GL072012.

Y. Liang, N. P. Gillett, and A. H. Monahan. Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend. *Geophysical Research Letters*, 47 (12), June 2020. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019GL086757. URL https://onlinelibrary.wiley.com/doi/10.1029/2019GL086757.

R. E. Lucas. Adjustment Costs and the Theory of Supply. *Journal of Political Economy*, 75(4): 321–334, 1967. ISSN 0022-3808. URL https://www.jstor.org/stable/1828594.

A. McDonnell. Model democracy. https://zenodo.org/doi/10.5281/zenodo.12629997, 2024. [CODE AND DATASET].

A. L. Merrifield, L. Brunner, R. Lorenz, V. Humphrey, and R. Knutti. Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications. *Geoscientific Model Development*, 16(16):4715–4747, Aug. 2023. ISSN 1991-9603. doi: 10.5194/gmd-16-4715-2023. URL https://gmd.copernicus.org/articles/16/4715/2023/.

M. Mussa. External and Internal Adjustment Costs and the Theory of Aggregate and Firm Investment. *Economica*, 44(174):163, May 1977. ISSN 00130427. doi: 10.2307/2553718. URL https://www.jstor.org/stable/10.2307/2553718?origin=crossref.

A. Ribes, S. Qasmi, and N. P. Gillett. Making climate projections conditional on historical observations. *Science Advances*, 7(4):eabc0671, Jan. 2021. ISSN 2375-2548. doi: 10.1126/sciadv.abc0671. URL https://www.science.org/doi/10.1126/sciadv.abc0671.

S. C. Sherwood, M. J. Webb, J. D. Annan, K. C. Armour, P. M. Forster, J. C. Hargreaves, G. Hegerl, S. A. Klein, K. D. Marvel, E. J. Rohling, et al. An assessment of earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, 58(4):e2019RG000678, 2020.

H. Shiogama, M. Watanabe, H. Kim, and N. Hirota. Emergent constraints on future precipitation changes. *Nature*, 602(7898):612–616, 2022.

K. Tokarska. Past warming trend constrains future warming in CMIP6 models. *Science Advances*, page 14, 2020.

M. D. Zelinka, T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor. Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, 47(1), Jan. 2020. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019GL085782. URL https://onlinelibrary.wiley.com/doi/10.1029/2019GL085782.

Supplementary Materials: To what extent does discounting 'hot' climate models improve the predictive skill of climate model ensembles?

Abigail McDonnell*[1], Adam Michael Bauer[2], and Cristian Proistosescu[1,3]

[1] *Department of Climate, Meteorology, and Atmospheric Sciences, University of Illinois Urbana-Champaign, 1301 W Green St, Urbana IL 61801*

[2] *Department of Physics, University of Illinois Urbana-Champaign, 1110 W Green St, Loomis Laboratory, Urbana, IL 61801*

[3] *Department of Earth Sciences and Environmental Change, University of Illinois Urbana-Champaign, Urbana, IL 61801*

Forthcoming in *Earth's Future*

September 20, 2024

Contents

*Corresponding author email: amm18@illinois.edu

S1 Methods

S1.1 Data

We use CMIP6 model historical data and SSP5–8.5 CMIP6 projections that include temperature and precipitation time series for the period of 1850-2099. The SSP5–8.5 scenario is used because it is expected to have the highest signal-to-noise ratio of forced response to internal variability. Predictive skill is expected to be worse in scenarios with weaker forcing due to the increased importance of unforced natural variability. The temperature ensemble consists of 27 members and the precipitation ensemble consists of 26 members. The data was downloaded from a public repository (https://esgf-data.dkrz.de/search/cmip6-dkrz/) and processed using the Pangeo stack and `Python` libraries such as `xarray` and `gcsfs`. We removed the seasonal cycle from the data and took the area-weighted global average over the 1850–2099 time period to compute the global average time series used in Figure 1 (see main text) and each calculation thereafter.

Table S1: CMIP6 data information.

Activity	Climate Model	Experiment	Member	Table
CMIP, ScenarioMIP	CanESM5	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	AWI-CM-1-1-MR	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	EC-Earth3-Veg	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	CMCC-CM2-SR5	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	TaiESM1	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	IPSL-CM6A-LR	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	NESM3	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	CAMS-CSM1-0	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	FGOALS-f3-L	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	GFDL-CM4	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	CMCC-ESM2	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	MRI-ESM2-0	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	INM-CM5-0	ssp585, historical	r1i1p1f1	Amon

Continued on next page

Table S1 – *Continued from previous page*

Activity	Climate Model	Experiment	Member	Table
CMIP, ScenarioMIP	BCC-CSM2-MR	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	INM-CM4-8	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	EC-Earth3	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	IITM-ESM	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	CAS-ESM2-0	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	CESM2-WACCM	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	EC-Earth3-CC	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	E3SM-1-1	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	GFDL-ESM4	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	MIROC6	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	FIO-ESM-2-0	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	EC-Earth3-Veg-LR	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	KACE-1-0-G	ssp585, historical	r1i1p1f1	Amon
CMIP, ScenarioMIP	NorESM2-MM	ssp585, historical	r1i1p1f1	Amon

S1.2 Relative forecast error

To quantify the degree to which applying a weighting scheme reduces variance in an ensemble projection, we introduce the relative forecast error (RFE), defined as

$$RFE := \frac{\sigma_w^2}{\sigma_{uw}^2} = \frac{\langle x^2\rangle_w - \langle x\rangle_w^2}{\langle x^2\rangle - \langle x\rangle^2}, \tag{0.1}$$

where σ_w^2 is the variance of the weighted projection and σ_{uw}^2 is the variance of the unweighted projection. An RFE of more than one implies the weighted projection is more uncertain than the unweighted projection; an RFE of less than one indicates that the spread in the weighted ensemble is less than that of the unweighted projection.

S1.3 Weighting scheme

S1.3.1 Theoretical framework

We follow the approach of Liang et al. (2020) and calculate a given model's weight by determining the distance between each ensemble member and a chosen 'pseudo-observation'. The pseudo-observation is treated as a "true" observation in the spirit of a perfect model test (Liang et al., 2020). The weighting scheme depends on two characteristics of the model ensemble: (i) the ability of a given model to reproduce the pseudo-observation, and (ii) model interdependence (see Knutti et al. (2017) for further discussion).

Formally, a set of ensemble weights can be computed using the following prescription. Consider a set of climate models $\mathcal{M}$, and allow each member $i \in \mathcal{M}$ to have a set of weighting metrics, $\mathcal{L}$, (i.e., temperature, precipitation, etc.) given by $\xi_i^{(\ell)} \in \mathcal{L}$, with trend over the historical period given by $\tilde{\xi}_{i,hist}^{(\ell)}$. Let the chosen pseudo-observation be indexed by $i^* \in \mathcal{M}$. Then the weight for a given model $i \in \mathcal{M}$ with chosen pseudo-observation $i^* \in \mathcal{M}$ is described by,

$$w_i^{(i^*)} = \frac{e^{-D_{i,i^*}^2/\sigma_D^2}}{1 + \sum_{j \neq i}^{M} e^{-S_{i,j}^2/\sigma_S^2}}, \tag{0.2}$$

where

$$D_{i,i^*} = \sqrt{\sum_{\ell \in \mathcal{L}} \left[\frac{\tilde{\xi}_{i,hist}^{(\ell)} - \tilde{\xi}_{i^*,hist}^{(\ell)}}{\mathrm{med}\left(S_{i,i^*}^{(\ell)}\right)} \right]^2} \tag{0.3}$$

is the normalized L^2-distance in trend-space between the weighting metric trend(s) of the chosen model and pseudo-observation for the historical period,

$$S_{i,j}^{(\ell)} = \sqrt{\sum_{\ell \in \mathcal{L}} \left[\frac{\tilde{\xi}_{i,hist}^{(\ell)} - \tilde{\xi}_{j,hist}^{(\ell)}}{\mathrm{med}\left(S_{i,j}^{(\ell)}\right)} \right]^2}, \tag{0.4}$$

is the L^2-distance between two models $i, j \in \mathcal{M}$ (with $i \neq j \neq i^*$) normalized by each median $\mathrm{med}\left(S_{i,j}^{(\ell)}\right)$ in trend-space, $M := |\mathcal{M}|$ is the number of models in the ensemble, and σ_D, σ_S are shape parameters. A smaller value of σ_D will assign a substantive amount of weight to a small number of models that are similar to pseudo-observation. Conversely, large values of σ_D is similar to an unweighted ensemble. σ_S functions similarly, but for model interdependence. In this case,

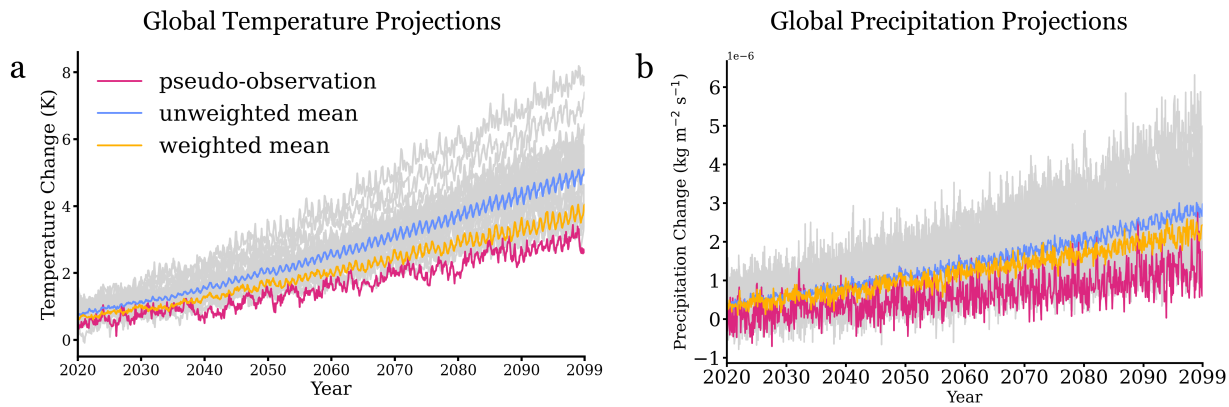

Figure S1: Weighting Scheme Example. Panels **a–b** show pseudo-observation impact on weighted temperature and precipitation projections. Panel **a** shows weighted (yellow line) and unweighted (blue line) mean temperature projections with weights based on historical temperature trends for the chosen pseudo-observation (pink line). The entire ensemble is shown in the grey lines. Panel **b** is as Panel **a**, but for precipitation projections. Note all panels show anomalies above the historical mean.

we have $\xi_i^{(\ell)} = T_i$, that is, we are only weighting by global temperature; likewise, $\mathcal{L} = \{T_i\}$.

To produce the distribution of perfect model test errors seen in Figure 1 (see main text), we choose a member of our climate model ensemble as the pseudo-observation. We then, using (0.2), compute the ensemble weights for that choice of pseudo-observation, and apply them to the remaining ensemble members to compute the weighted mean. An example of this being applied for a single pseudo-observation choice is shown in Figure S1. The perfect model test error for the chosen pseudo-observation is the difference between end-of-century warming relative to the historical period summarized in Table S2 for the pseudo-observation and weighted mean projection. Carrying out this process recursively, where each climate model ensemble member is chosen as the pseudo-observation once, produces the distribution of perfect model test errors seen in Figure 1. The same procedure using the unweighted mean for every choice of pseudo-observation gives the distribution of perfect model test errors for the unweighted mean. The RMSE reduction is found by comparing the RMSE of these two distributions, using the usual definition for RMSE.

S1.3.2 Calibration of weighting scheme

We define the historical period as $X - 2014$ where X is the year with maximum correlation between the historical trend of the weighting metric (i.e., historical global average temperature) and future

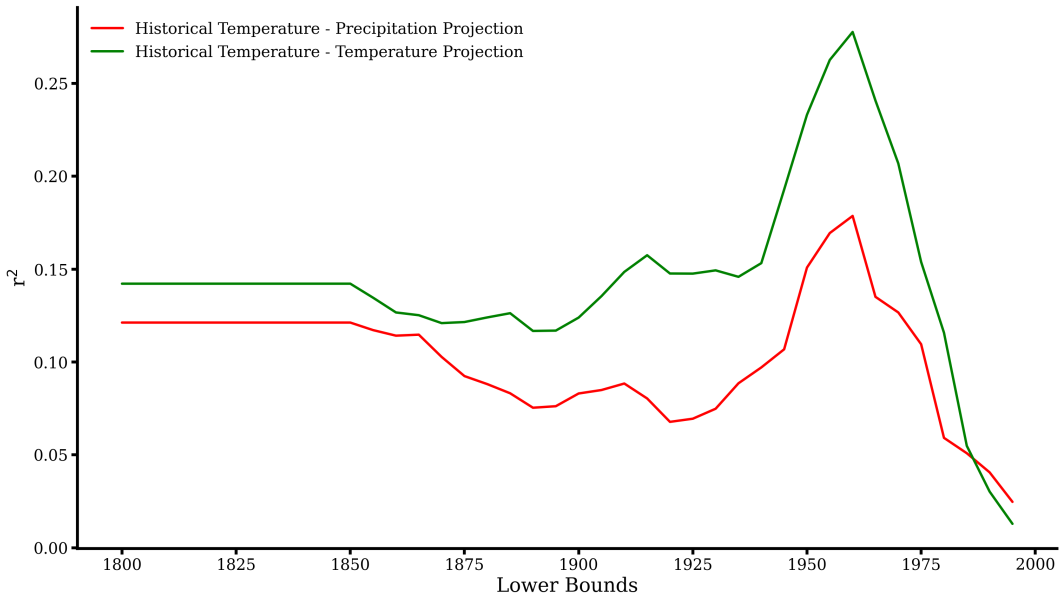

Figure S2: Correlation between historical and future trends for different definitions of the historical period. Shown is the correlation coefficient (r^2) between historical temperature trends and future temperature trends (green line); historical temperature trends and future precipitation trends (red line) as a function of what lower bound is chosen for the historical period; i.e., for different choices of X in the time period $X - 2014$.

Table S2: Defining the historical period. Listed is the lower bound X of the historical period $X - 2014$ for each climate variable and weighting metric. The historical lower bound is the value along the abscissa of Figure S2 where the maximum correlation between climate variable and weighting metric occurs.

Climate variable	Weighting metric	Historical lower bound
Temperature	Temperature	1960
Precipitation	Temperature	1960

trends for the variable being weighted (i.e., future temperature projections) (see Figure S2). Our results are summarized in Table S2 for each climate variable and weighting metric.

To calibrate the optimal values for the shape parameters σ_D and σ_S, we perform a gridded optimization routine inspired by the calibration scheme in Knutti et al. (2017) (see their Figure 3c) that we outline in Algorithm S1. The idea is to choose some threshold for the fraction of pseudo-observations that should fall within the predicted range of our weighted mean, which we call F^*. Then for each possible pair of (σ_D, σ_S), we cycle through choices of pseudo-observation, compute the weighted mean and corresponding standard deviation of the prediction, and ask if the pseudo-observation lies within the $\pm 2\sigma$ predicted range. This approach is repeated for the entire

(σ_D, σ_S) grid.

The result of this procedure is shown in Figure S3, where we show the percentage of pseudo-observations that lie within the predicted range on our (σ_D, σ_S) grid. We then choose our optimal shape parameters (σ_D^*, σ_S^*) by minimizing the L^2-norm subject to the fraction of pseudo-observations that were predicted are above F^*. We choose $F^* \sim .89$ for our temperature-based weights.

Algorithm S1 Calibration of shape parameters σ_D and σ_S.

Data: A set of $\mathcal{M}$ climate models, where $M := |\mathcal{M}|$ is the number of models in the ensemble.

Initialize: Choose a climate variable of interest (i.e., temperature) given by ζ and weighting metric (i.e., historical temperature) given by ξ. Define the historical period and future period. Remove seasonal cycle and take area-weighted global average of both the climate variable and weighting metric. Compute trends over historical period; define $\tilde{\xi}_{i,hist}$ as the historical trend in the weighting metric for model $i \in \mathcal{M}$. Define $\bar{\zeta}_{i,fut}$ as the average of the climate variable over the future period for model $i \in \mathcal{M}$. Choose a $\bar{\sigma}_D$ and $\bar{\sigma}_S$, and define $\Sigma_D := [0, \bar{\sigma}_D]$ and $\Sigma_S := [0, \bar{\sigma}_S]$ with coarseness $\theta > 0$. Then $\Sigma_D \times \Sigma_S$ creates a grid of possible (σ_D, σ_S) pairs with θ^2 total candidates. Choose a minimum fraction of pseudo-observations that must be predicted by ensemble weighting, given by F^*.

Ensure: $\bar{\sigma}_D, \bar{\sigma}_S, \theta > 0$, $0 \leq F^* \leq 1$

for all $(\sigma_D, \sigma_S) \in \Sigma_D \times \Sigma_S$ **do**
  $N_{i^*} \leftarrow 0$ ▷ Initialize number of pseudo-observations in predicted range
  for all $i^* \in \mathcal{M}$ **do** ▷ Iterate through each choice of pseudo-observation
    $\langle\zeta\rangle_w \leftarrow \sum_{i \in \mathcal{M}\setminus\{i^*\}} w_i^{(i^*)} \bar{\zeta}_{i,fut}$ ▷ Using Eqn. (0.2) for $w_i^{(i^*)}$
    $\langle\zeta^2\rangle_w \leftarrow \sum_{i \in \mathcal{M}\setminus\{i^*\}} w_i^{(i^*)} \bar{\zeta}_{i,fut}^2$
    $\sigma_\zeta \leftarrow \sqrt{\langle\zeta^2\rangle_w - \langle\zeta\rangle_w^2}$
    if $\langle\zeta\rangle_w - 2\sigma_\zeta \leq \bar{\zeta}_{i^*,fut} \leq \langle\zeta\rangle_w + 2\sigma_\zeta$ **then**
      $N_{i^*} \leftarrow N_{i^*} + 1$ ▷ Increment N_{i^*}
    end if
  end for
  $F_{\sigma_D,\sigma_S} \leftarrow N_{i^*}/M$ ▷ Translate number to fraction
end for
$\Gamma^* \leftarrow \{(\sigma_D, \sigma_S) \in \Sigma_D \times \Sigma_S : F_{\sigma_D,\sigma_S} \geq F^*\}$ ▷ Region of candidate optimal (σ_D, σ_S) pairs
$(\sigma_D^*, \sigma_S^*) \leftarrow \arg\min_{(\sigma_D,\sigma_S) \in \Gamma^*} \sqrt{\sigma_D^2 + \sigma_S^2}$ ▷ Minimize L^2-norm for (σ_D, σ_S) pairs on Γ

Output: σ_D^*, σ_S^* ▷ Optimal values of σ_D and σ_S

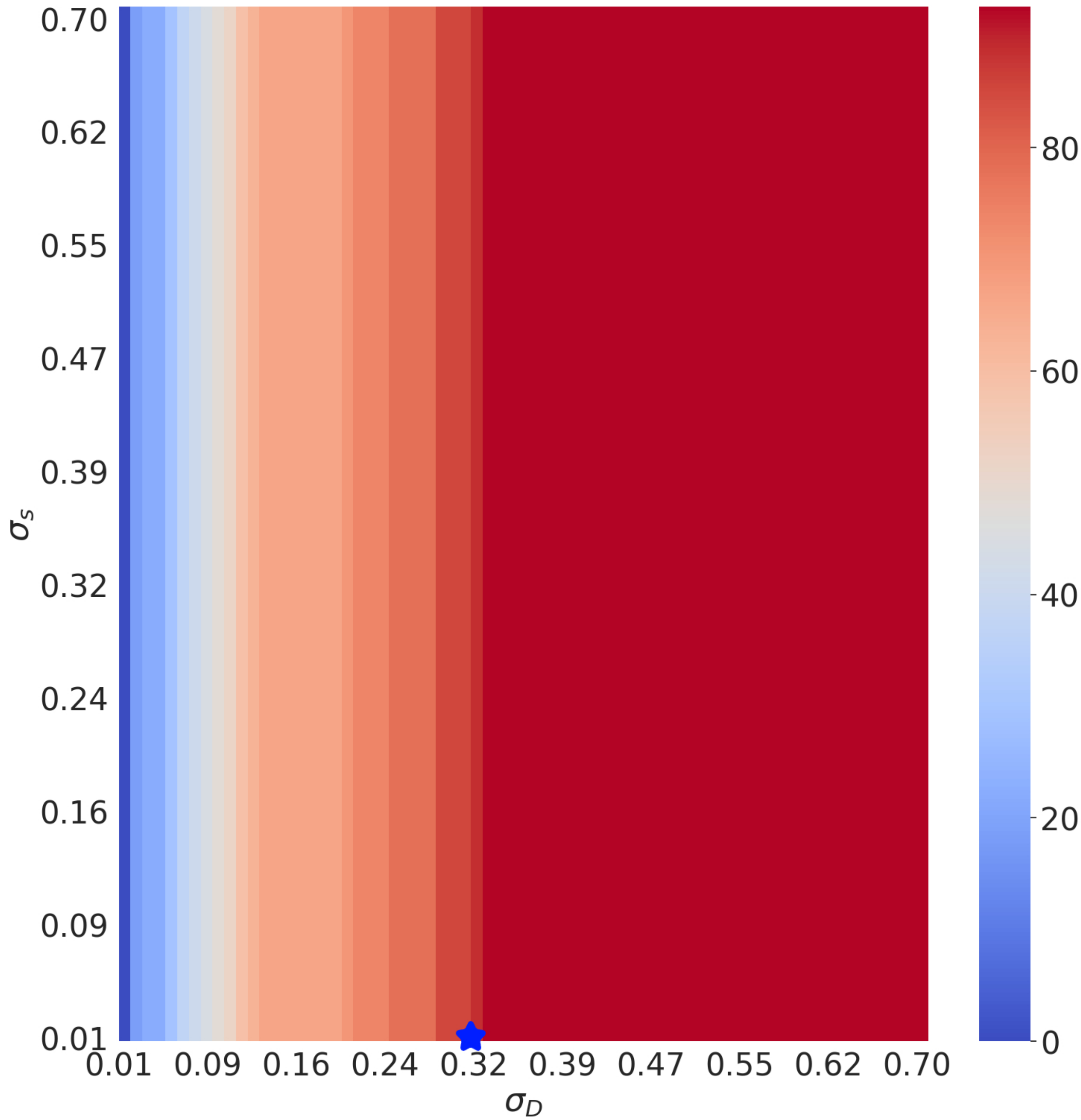

Figure S3: Heatmap of shape parameter calibration for temperature-based weights. Shown is percentage of pseudo-observations that are predicted in our weighted mean approach for each σ_D and σ_S combination. The blue star shows the optimal values of σ_D and σ_S.

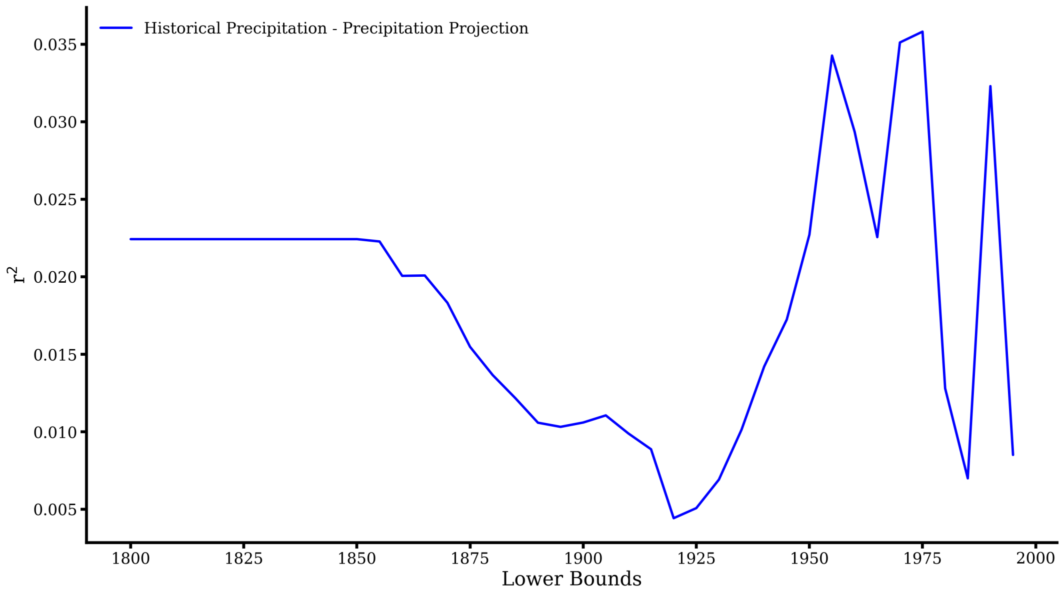

Figure S4: Correlation between historical and future trends for different definitions of the historical period. Shown is the correlation coefficient (r^2) between historical precipitation trends and future precipitation trends as a function of what lower bound is chosen for the historical period; i.e., for different choices of X in the time period $X - 2014$.

S2 Precipitation-based analog to IPCC's temperature-based weighting scheme

We here show the results of using a precipitation-based analog of our temperature-based weighting scheme for weighting future global and regional precipitation projections. To be clear, this scheme follows the same logic as our temperature-based scheme, but rather than weighting models by their ability to reproduce historical temperature trends, we use historical precipitation trends. The calibration scheme is exactly the same as we laid out in the *Methods* section above (see Figures S4 and S5).

For our global analysis, we find that global average precipitation RMSE is reduced by 9.23% using precipitation-based weights. See Figure S6 for the histogram of errors using this scheme. This implies that historical precipitation trends are a skillful predictor of future global precipitation trends, though less skillful than historical temperature trends (see the main text).

For the regional analysis, we find that there exists a heterogeneous pattern of RMSE reduction, similar to the temperature-based weights in the main text (Figure S7). We again find that this owes to the degree of correlation between global historical precipitation trends and regional precipitation

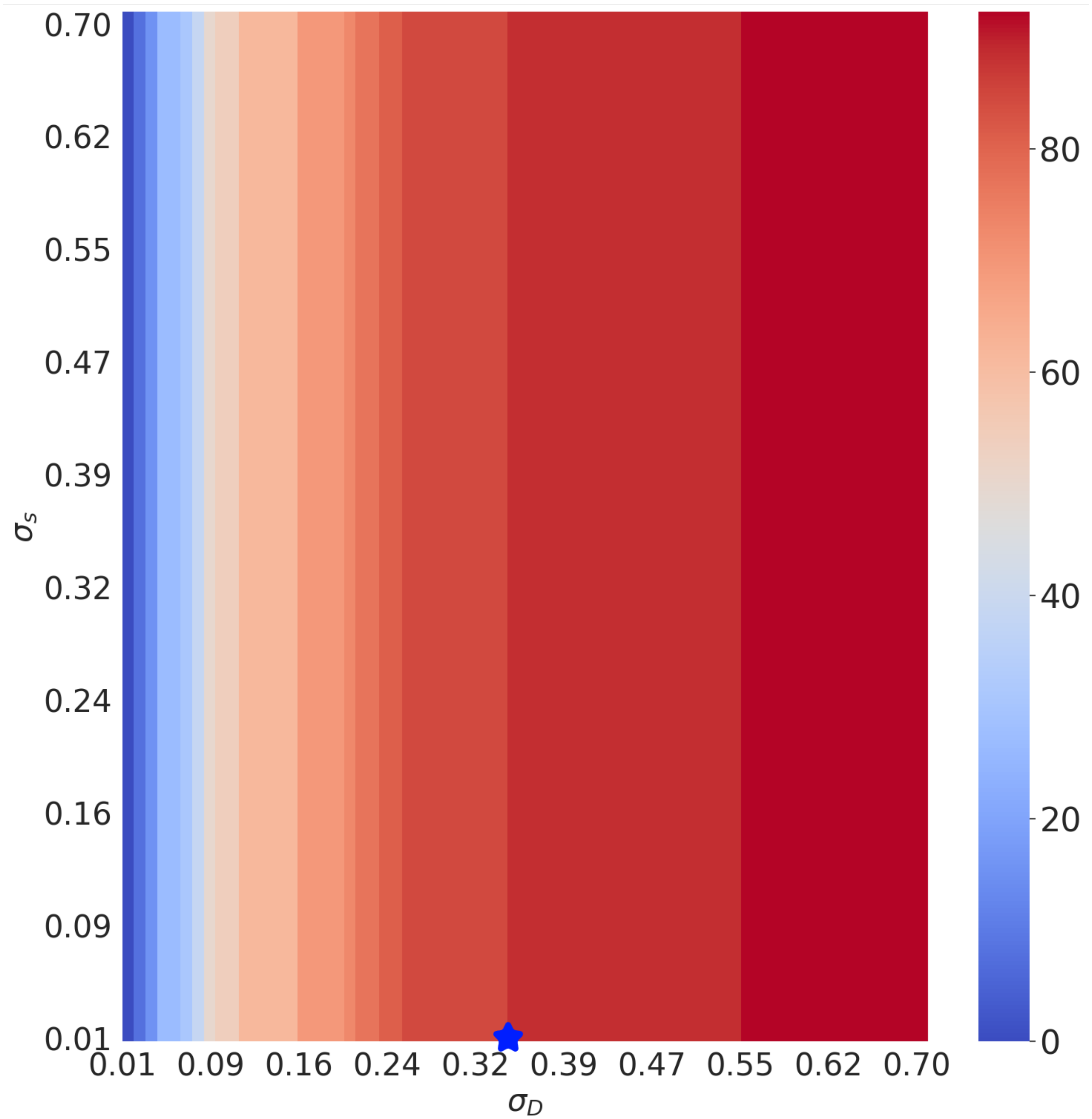

Figure S5: Heatmap of shape parameter calibration for precipitation-based weights. Shown is percentage of pseudo-observations that are predicted in our weighted mean approach for each σ_D and σ_S combination. The blue star shows the optimal values of σ_D and σ_S.

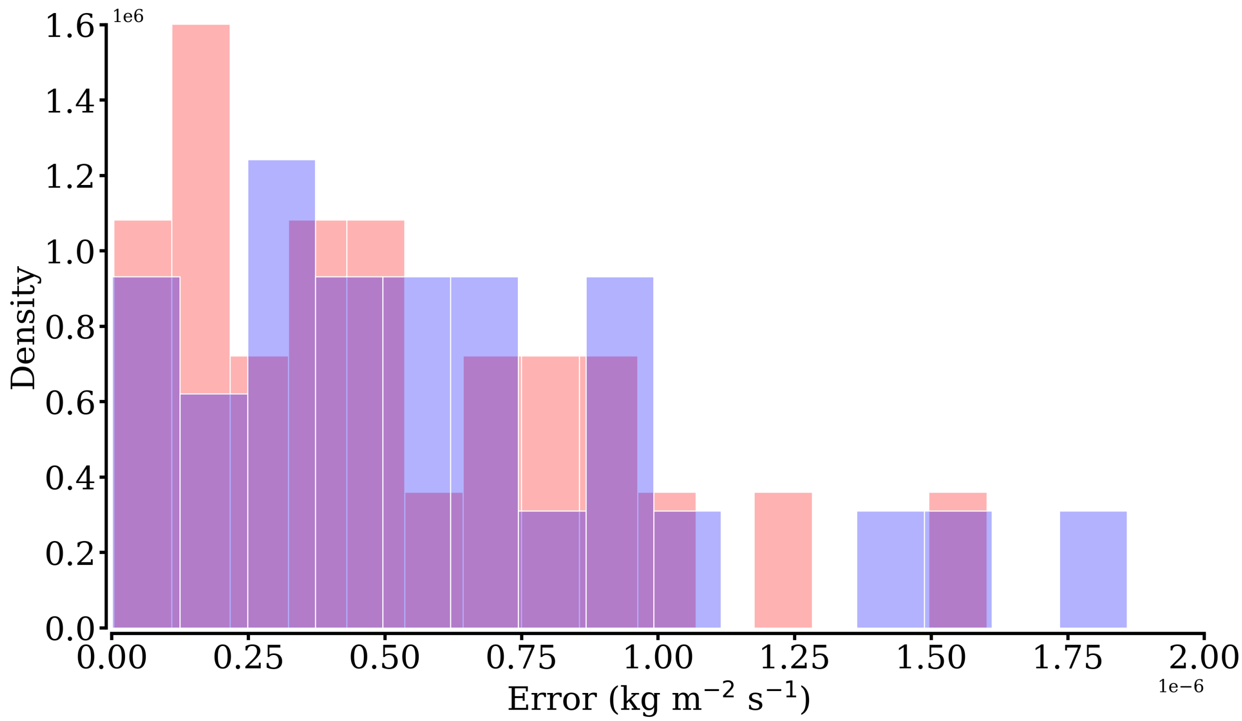

Figure S6: Perfect model test error distributions for global precipitation using historical precipitation-based weights.

projections varying substantially across space. Importantly, we find that there is a lack of precision gains in the low- and mid-latitudes, making global historical precipitation trends a poor metric for model weighting in regional impact analysis.

Table S3: Mean RMSE Change by Region

Country	**Temperature**	**Precipitation**
Afghanistan	0.74	1.12
Australia	0.89	1.05
Brazil	0.93	1.14
Chile	0.81	1.11
China	0.85	1.07
Democratic Republic of Congo	1.02	0.94
Greenland	0.83	0.95

Continued on next page

Table S3 – *Continued from previous page*

Country	**Temperature**	**Precipitation**
India	0.91	1.04
Indonesia	0.76	1.11
Mexico	0.80	1.12
Nepal	0.89	1.09
Russia	0.73	0.89
South Africa	0.76	1.07
Turkey	0.78	0.97
United Kingdom	0.91	1.08
United States of America	0.80	1.08

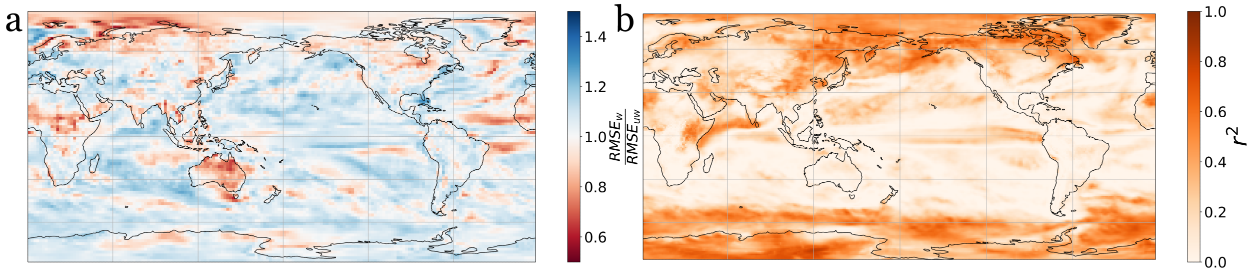

Figure S7: Regional decomposition of RMSE reduction and variance explained by weighting future precipitation projections using precipitation-based weights. Panel **a** shows the spatial distribution of the relative RMSE between the raw ensemble mean and our historical precipitation trend-derived weighting technique. Here, a value less than (greater than, resp.) unity implies a more (less, resp.) precise prediction using the weighted ensemble mean as opposed to the unweighted ensemble mean. Panel **b** shows the variance in future regional precipitation anomalies that is explained by historical global precipitation trends. Note that high levels of variance explained should correspond to a relative RMSE of less than one.

References

R. Knutti, J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring. A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, 2017. ISSN 00948276. doi: 10.1002/2016GL072012. URL http://doi.wiley.com/10.1002/2016GL072012.

Y. Liang, N. P. Gillett, and A. H. Monahan. Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend. *Geophysical Research Letters*, 47 (12), June 2020. ISSN 0094-8276, 1944-8007. doi: 10.1029/2019GL086757. URL https://onlinelibrary.wiley.com/doi/10.1029/2019GL086757.