ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™

LETTER • OPEN ACCESS

Revisiting two-layer energy balance models for climate assessment

To cite this article: Junichi Tsutsui and Chris Smith 2025 *Environ. Res. Lett.* **20** 014059

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Revisiting two-layer energy balance models for climate assessment

Junichi Tsutsui[1,*] and Chris Smith[2,3]

[1] Sustainable System Research Laboratory, Central Research Institute of Electric Power Industry, Abiko, Japan
[2] School of Earth and Environment, University of Leeds, Leeds, United Kingdom
[3] Energy, Climate and Environment Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
* Author to whom any correspondence should be addressed.

E-mail: tsutsui@criepi.denken.or.jp

Keywords: emulator, CO_2 forcing, probabilistic climate assessment, constraining, Metropolis-Hastings sampler

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
25 July 2024

REVISED
18 October 2024

ACCEPTED FOR PUBLICATION
13 December 2024

PUBLISHED
27 December 2024

Abstract

Given the pivotal role of probabilistic approaches with two-layer energy balance models in the latest climate assessment, this study aims to gain deeper insight into their advancement by comparing different approaches for generating constrained posterior ensembles. Several methodological improvements are possible both in the calibration of model parameters to the behavior of comprehensive Earth system models and in constraining the calibrated parameter ensemble with other lines of evidence. The results imply that a conventional single parameter representing evolving climate feedback characteristics is not a requirement for reliable climate projections; rather, there are potential improvements on the forcing side regarding the separation of forcing and feedbacks. Constraining the ensemble based on observational and expert-assessed climate metrics, which critically affects probabilistic climate assessment, needs to appropriately deal with different constraints on a multivariate space in a standardized and flexible way. The method introduced here is an option that fulfills the need.

1. Introduction

In the sixth assessment of the Intergovernmental Panel on Climate Change (IPCC AR6) Working Group I (WGI) contribution, the two-layer model shown in figure 1 was intensively used to assess global warming levels and associated thermosteric sea-level changes in response to changes in CO_2 and other forcing agents [1–3]. The model formulation is the one proposed in [4], and termed as EBM-ε [5], energy balance model with an efficacy factor (ε) for deep-ocean heat uptake [6]. Despite the difficulty of dealing with the evolving heat uptake characteristics [7], a two-layer model is simple and sufficiently accurate to emulate complex full-scale Earth system models (ESMs) in terms of the relationship between idealized CO_2 forcing and global annual mean temperature response [5, 8]. In fact, ESM simulations are the basis for EBM-ε and for other reduced complexity models, or emulators [9], and emulator-based assessment in AR6 was based on the comprehensive ESM dataset provided by the Coupled Model Intercomparison Project Phase 6 (CMIP6) [10].

Simplicity is crucial for probabilistic climate assessment [11–13] using a large-member parameter ensemble that requires time efficiency in production runs and conforms to multiple lines of evidence in terms of forcing-response characteristics. The AR6 approach, described in AR6 WGI Chapter 7 [1, 14], uses a final posterior set of approximately 2,000 members, constrained from a one million-member prior. This constrained ensemble was applied to time integrations ranging from hundreds to thousands of years over numerous conditions to evaluate the warming contributions of different forcing agents, including greenhouse gases and aerosols.

An ensemble for an emulator was generated through calibration and sampling processes [15]. Calibration is a numerical optimization for a set of parameters to minimize differences between an

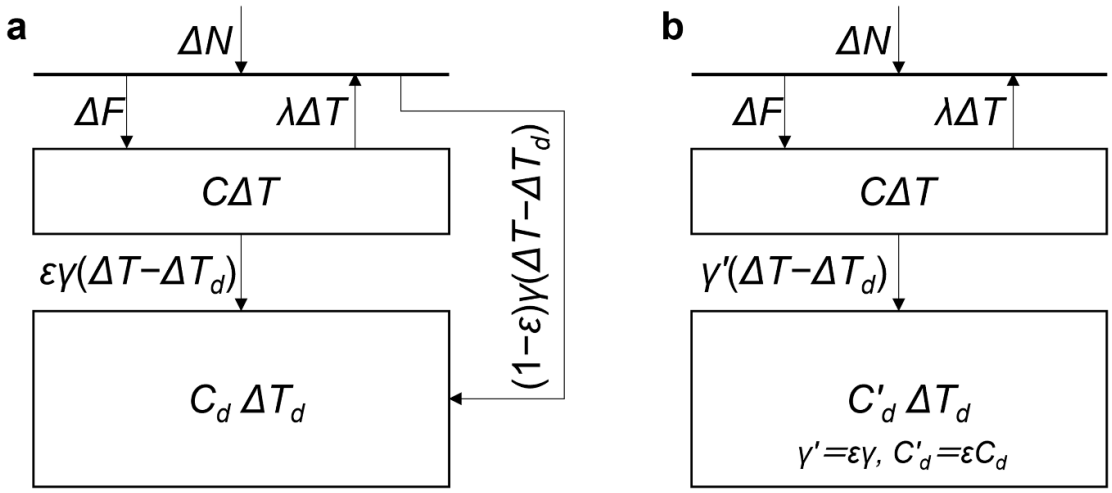

Figure 1. Schematic diagram of the two-layer energy-balance model with the ocean heat uptake efficacy. The arrows represent energy flow between the surface and deep ocean layers, and the top of the atmosphere (the top bar). Here, ΔN is the net energy flux, ΔF is the forcing change, ΔT and ΔT_d are the surface and deep ocean temperature changes, C and C_d are the heat capacities of the surface and deep ocean layers, λ is the climate feedback parameter, γ is the heat transfer coefficient between the two layers, and ε is the efficacy of deep ocean heat uptake. (a) Original form, where bypassing arrow on the right represents assumed energy flow associated with the heat uptake efficacy, illustrated in the same way as in [23]. (b) Equivalent form, where the efficacy-assuming flow is implicitly embedded in the deep layer with $\gamma' = \varepsilon\gamma$ and $C'_d = \varepsilon C_d$. The two forms are mathematically equivalent, as ΔN in the latter form is diagnostically adjusted. See supplementary 1.1 for comprehensive descriptions.

emulator and an ESM, which will usually be performed for all available ESMs and for several variables. Sampling typically consists of two stages: random sampling according to the probability density of the calibrated ensemble; and constraining the range of the sampled ensemble to fit evidence other than the base ESMs, such as observations and process understanding. In the AR6 approach, a parameter ensemble for the two-layer model that describes the climate response was combined with a forcing ensemble that represents the uncertainties across different forcing agents. The result of these processes is the synthesis of multiple lines of evidence, and the application of the combined climate-response/forcing ensemble to many emissions scenarios [16–18], providing an emulator can also translate emissions to forcing [19], allows for knowledge transfer from climate science to climate change mitigation [20]. This synthesis is the most essential role of simple emulators.

Besides the probabilistic approach, there is also a method of mapping EBM-ε parameters to specific values of two climate sensitivity metrics—equilibrium climate sensitivity (ECS) and transient climate response (TCR)—which was applied in AR6 WGI Chapter 4 and 9 [2, 3] to assess future global climate under illustrative scenarios. In this approach, climate projections with EBM-ε configured with AR6-assessed ranges of ECS and TCR were used in conjunction with results from comprehensive CMIP6 scenario experiments [21].

Considering the importance of EBM-ε, the present study explores different methods to advance probabilistic approaches by comparing the AR6 method with alternative implementations. One previous study [22] examined the performance of EBM-ε and other models and presented emulation errors and potential improvements in calibration processes. Some of the findings and implications are pertinent to the results of the present study and are addressed in the discussion section.

2. Method

Table 1 summarizes four experimental cases, including the AR6 Chapter 7-equivalent of Case 0. Cases 1–3 are designed to compare several modifications regarding different calibrations, treatment of the ocean heat uptake efficacy, fidelity to ESMs for forcing-response properties, treatment of multivariate constraints, and scaling of CO_2 forcing. These modifications are specific to either process of calibration and sampling/constraining and will be described in the relevant result subsections.

The sampling process statistically generates random variable series from distributions informed by the calibrations to individual ESMs in a way that maintains the multi-ESM variance-covariance structure in a multivariate parameter space. The random series are constrained such that the ranges of several climate indicators from historical emulation runs match the ranges from their observation-based reference as closely as possible. In this study, the historical runs were conducted with the perturbed forcing ensemble used in AR6 WGI Chapter 7. The constraints were designed to reflect the observed changes in global surface air temperature (GSAT) in a recent past period of 1995–2014 and the observed ocean heat uptake from 1971 to 2018, which are

Table 1. Different cases for generating a large-member parameter ensemble.

#	Label	Calibration	Sampling	Constraining	CO_2 forcing for emulation runs
0	EBM-ε AR6 orig	WGI Chapter 7	q_{4x} and λ replaced independently	Intersection of individual constraint ranges	Perturbed with AR6-assessed q_{2x}
1	EBM-ε AR6		No replacing	Acceptance-rejection following probability density based on constraint ranges	Perturbed as in #0 and scaled by calibrated q_{4x}
2	EBM-ε S21	Reference [25]			
3	MCE-2l (standard EBM)	Reference [26]			Perturbed as in #0 and scaled by calibrated q_{4x} and an additional amplification factor

MCE-2l stands for a two-layer version of the Minimal CMIP Emulator [27], an implementation based on standard EBM, or a pure impulse response model, combined with its own CO_2 scaling scheme. q_{4x} and q_{2x} are forcing levels of quadrupling and doubling CO_2. See supplementary 1.1 for model formulations, and supplementary 1.2–1.3 for differences about calibration/sampling/constraining procedures and relevant CO_2 forcing schemes.

a subset of those used in AR6 WGI Chapter 7. These indicators were implemented in the alternative cases with a Metropolis-Hastings (MH) independent sampler [24], a general acceptance/rejection method following probability density based on constraint ranges.

To examine the differences in future climate projections between the constrained parameter ensembles developed here, emulation runs were conducted for future climate projections with five illustrative scenarios ranging from low to high emissions, as in AR6 WGI Chapter 4 [2], for which perturbed forcing series were used as in the historical runs.

See Supplementary Information for further details, as noted in table 1. The following sections describe and discuss the accuracy of individual calibrations, coverage of multiple ESMs, and consistency with other evidence.

3. Results

3.1. Calibration to individual ESMs

Typically, calibration is conducted for ESM output of top-of-the-atmosphere (TOA) net energy flux (ΔN) and the surface temperature change (ΔT) over 150 years after instantaneous quadrupling of the atmospheric CO_2 concentration [28]. This is also for the cases in table 1, except that Case 3 additionally includes ESM output from a transient CO_2 increase experiment, where the concentration increases at a 1% per year until it quadruples after 140 years. Here, the results are compared between EBM-ε with two different calibrated parameters [1, 25] (Cases 1 and 2; Case 0 is identical to Case 1) and a standard form of EBM, i.e. $\varepsilon = 1$, combined with a super-logarithmic CO_2 radiative forcing scaling [26] (Case 3). The comparison focuses on how accurately the behavior of different ESMs can be emulated for time series of ΔN and ΔT and their relations from the two experiments.

Figure 2 compares the results for a specific ESM as a typical case, which is one of those shown in supplementary table 1 and supplementary figure 1 for each of 46 ESMs used in this study. Although all emulated time series generally represent long-term ESM tendencies well, there are some methodological differences regarding (1) the scaling of transient CO_2 forcing changes and (2) ocean heat uptake efficacy.

The first point is found from a better performance of Case 3 in emulating ΔT in the transient experiment (blue marker and line in figure 2(b)). This performance can be measured by comparing the TCR between the ESM output and the derivation from the EBM parameters (supplementary figure 2). Forcing changes to the first and second doubling of CO_2 concentrations in the 1%-per-year increase trajectory are not necessarily the same [30] or scaled with a fixed factor, unlike the use of a standard approximate log-concentration formula [14, 31]. Case 3 incorporated an amplification factor from the first to the second doubling, thereby resulting in the successful emulation of all ESMs, as previously examined [26]. This elaborate scaling would be beneficial for the other cases to better emulate some CMIP6 models with greater amplification.

The second point is associated with different curvature of (ΔN, ΔT) trajectories in response to instantaneous CO_2 quadrupling (red marker and line in figure 2(c)), for which the ocean heat uptake efficacy in EBM-ε is responsible. As expected, the emulated trajectories were curved in Cases 1 and 2 and linear in Case 3. The degree of curvature is not necessarily the same in the former cases and depends on the different optimizations in their calibration.

In fact, ESM response to step forcing like a quadrupling CO_2 increase often shows a concave trajectory on a ΔN–ΔT plane, such that initial ΔN and eventual ΔT are both shifted from an assumed linear line toward larger values [32, 33]. The intercept points on the y- and x-axis correspond to the level

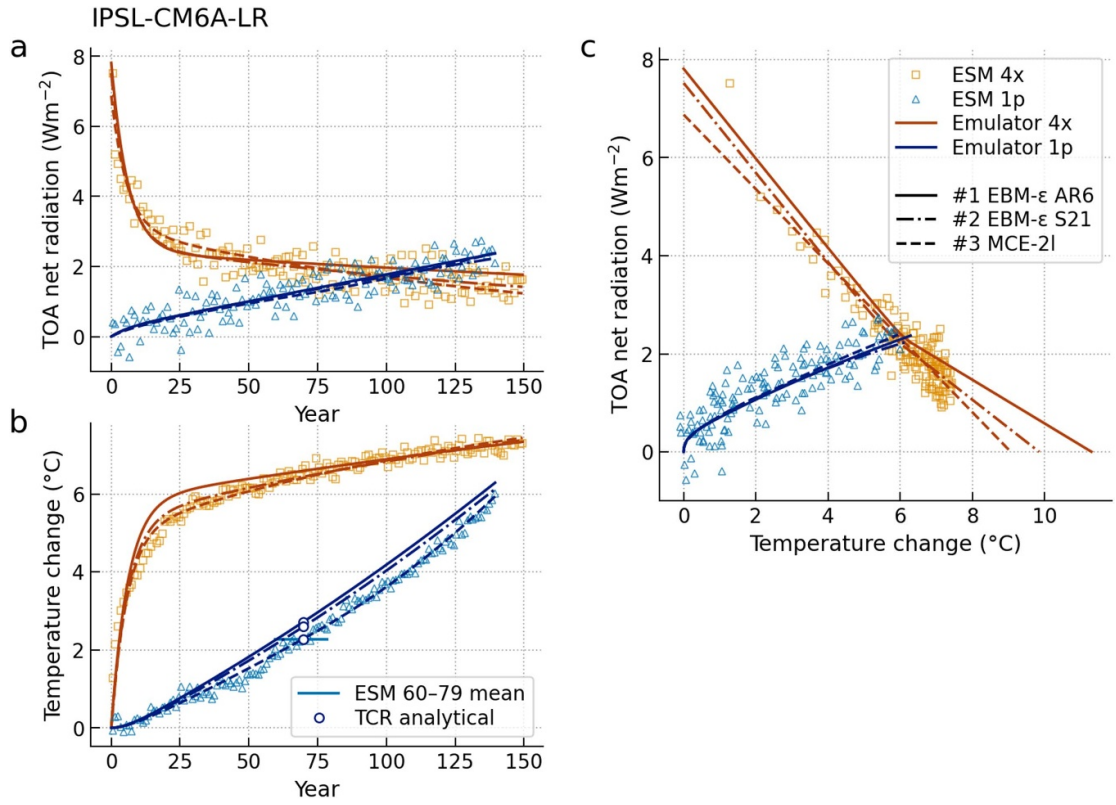

Figure 2. Emulation of two idealized CO_2 increase experiments with an individual ESM. As an example of ESM, IPSL-CM6A-LR [29] is selected. See supplementary figure 1 panels S1–S46 for all the ESMs used in this study. The CO_2 increase experiments are 150-year abrupt-4xCO2 and 140-year 1pctCO2 [10], labeled in the legend as '4x' and '1p', respectively. The markers show annual-mean ESM output, and the lines show emulation results with the three different methods. (a) Timeseries of TOA net energy flux. (b) Timeseries of changes in the surface temperature. (c) Relationship between the two variables. The emulation period of abrupt-4xCO2 is extended to year 3000.

of quadrupling CO_2 forcing (q_{4x}) and an equilibrium temperature change under q_{4x}; scaling them down to the level of doubling CO_2 (q_{2x}) provides an estimate of the ECS. This is a well-established method for estimating the ECS from an ESM experiment, as given in [34], and a curved trajectory may provide a better estimate than a linear one [35]. However, there are several CMIP6 models indicating relatively large differences in the intercept points between Cases 1 and 2, thereby implying that the ECS estimation is sensitive to calibration procedures.

In addition to climate sensitivity issues, incorporating the efficacy factor did not significantly affect the emulation accuracy of at most a 150 year time series from the two idealized experiments. These findings basically agree with the results of the Reduced Complexity Model Intercomparison Project (RCMIP) Phase 1 [9].

3.2. Sampling and constraining

Hereafter, a set of parameter ensembles and their derivatives are referred to as ESM-calibrated, unconstrained, and constrained ensembles, which correspond to the outputs of the calibration, sampling, and constraining, respectively.

In the sampling stage, modifications in Cases 1–3 affect statistics relevant to the climate feedback parameter (λ) and forcing from a doubling CO_2 (q_{2x}). The original Case 0 replaces these two components of the unconstrained ensemble with an independently generated series based on the AR6-assessed ranges [1]. Given their dominant role in the temperature response, this replacement ensures that the unconstrained ensemble is consistent with the AR6 assessment. Unlike the original, the alternatives leave their unconstrained ensemble until the constraining stage to represent the statistics of their ESM-calibrated ensemble as closely as possible.

As expected, by using ESM-derived calibrations, the relationship between q_{2x} and λ is maintained. Figure 3 indicates a weak positive correlation in all the calibrated ensembles inherited through the sampling and constraining stages (figures 3(b)–(d)), except for Case 0 (figure 3(a)). λ also has weak correlations with the other parameters although not as distinct as with q_{2x} (supplementary table 2).

Figure 4 compares the ranges of five key indicators: q_{2x}, λ, ECS, TCR, and the ratio of TCR to ECS, where the two climate sensitivity metrics were derived from the model parameters (supplementary 1.4). The ratio of TCR to ECS is denoted as RWF70, which represents the realized warming fraction (RWF) [36] when the CO_2 concentration doubles in the 70th

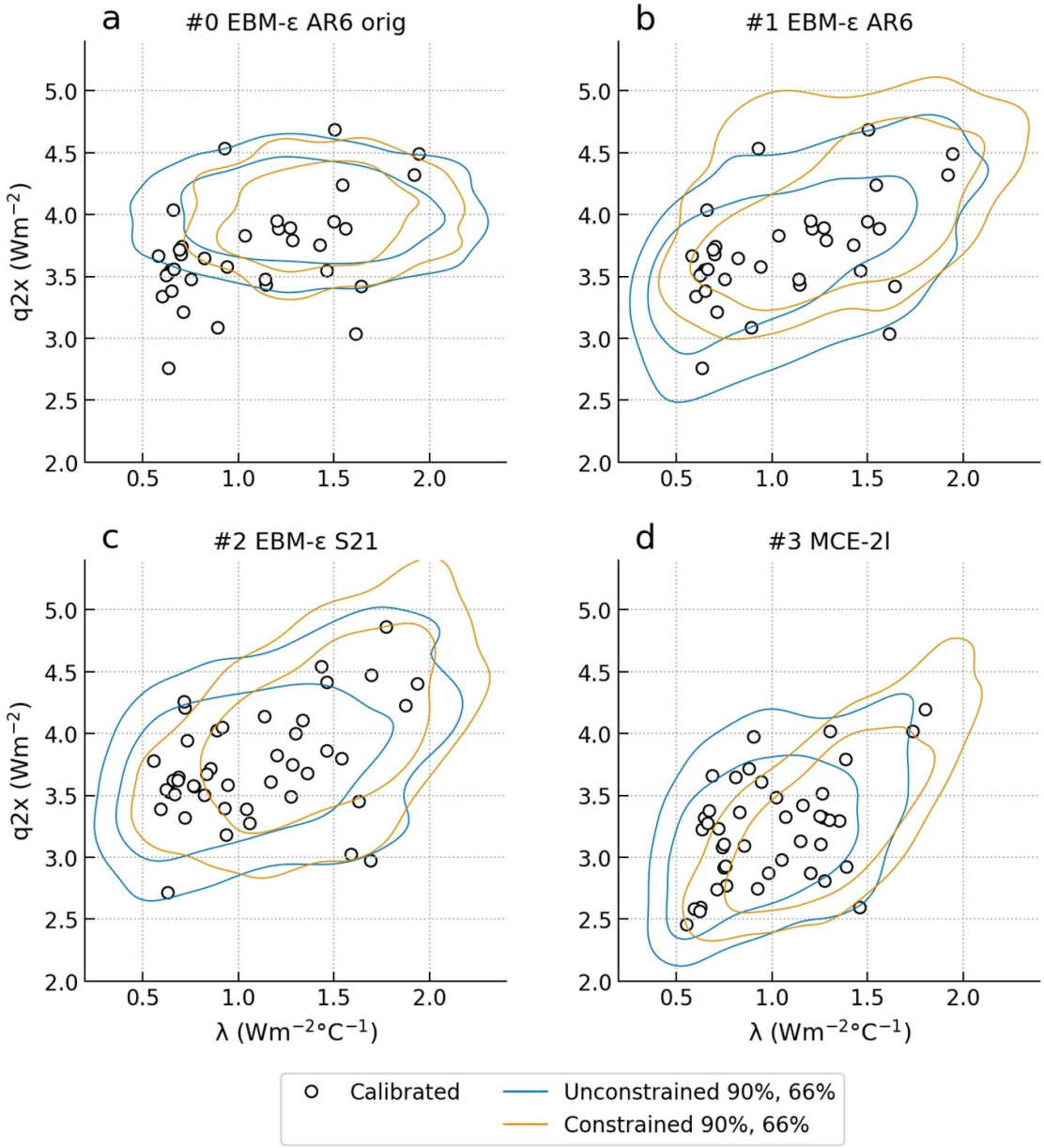

Figure 3. Relationships between the doubling CO_2 forcing and climate feedback parameter from calibrated, unconstrained, and constrained ensembles. The markers indicate individual calibrated values, and the contours indicate ranges covering 66% and 90% of each of the unconstrained and constrained ensembles.

year in a 1% per year increase trajectory. The ranges from the unconstrained ensembles widely cover those from the ESM-calibrated ensembles, except for Case 0 where q_{2x} and ECS, among others, are adjusted to bring their medians close to the AR6-assessed 3.93 Wm^{-2} and 3 °C. These medians reflect the replaced λ with a median of about 1.3 Wm^{-2} °C^{-1}.

Case 3 is distinguished from the others by a lower q_{2x} such that most of the *likely* (17%–83%) ranges do not overlap. Meanwhile, the ranges of climate sensitivity metrics, which are proportional to the ratio of q_{2x} to λ, are rather comparable between the cases. This is due to the large uncertainty in λ itself and its value that is also relatively smaller in Case 3. Differences between the cases were also observed in the range of extreme values. Case 3 is again distinguished by overall smaller *very likely* (5%–95%) ranges, reflecting fewer uncertainties in the calibration.

Overall, the constraining process works toward reducing climate sensitivity, and the constrained ranges are somewhat biased toward the lower side compared with the AR6-assessed ranges for both the ECS and TCR. Although most of the ranges overlap, the low-sensitivity bias was relatively large for Case 3. In contrast, the ranges of the GSAT and ocean heat uptake indicators from the historical emulation runs were similar for the four cases (supplementary figures 3 and 4). The characteristics of Case 3 are related to its relatively short response time scales, as observed from the greater RWF shown in figure 4(e).

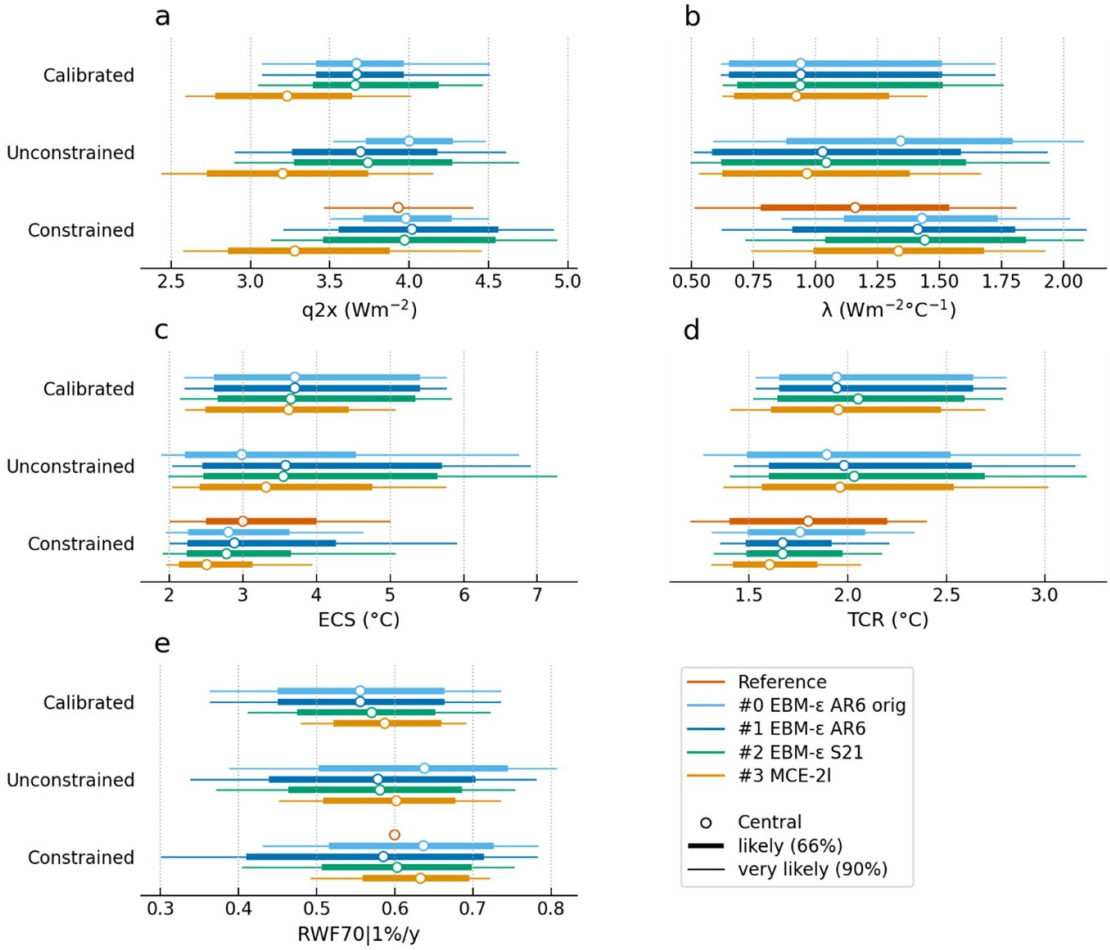

Figure 4. Ranges of selected key parameters from different ensembles denoted by colors, indicating *likely* (17%–83%) and *very likely* (5%–95%) ranges with a central value. (a) Doubling CO_2 forcing (q_{2x}); (b) Climate feedback parameter (λ); (c) Equilibrium climate sensitivity (ECS); (d) Transient climate response (TCR); (e) Realized warming fraction at doubling CO_2 in a 1%-per-year increase trajectory (TCR-to-ECS ratio). Each panel contains three grouped ranges: calibrated, unconstrained, and constrained. As a reference, the groups of the constrained ranges are accompanied with corresponding AR6-assessed ranges, based on [1], Table 7.4, 7.10, 7.13, and 7.14 for q_{2x}, λ, ECS, and TCR, respectively, and the ratio of the assessed central TCR to the assessed central ECS.

3.3. Future climate projections by constrained runs

Figure 5 compares the GSAT changes from the four ensembles for the mid-term (2041–2060) and long-term (2081–2100) 20 year periods relative to 1995–2014. The results for the four cases agreed with each other in terms of ensemble ranges as well as time series (supplementary figure 5), suggesting that the four ensembles would be nearly equivalent in terms of probabilistic climate projections. The small differences between cases confirm that the MH sampler introduced in the alternative cases is working properly. See supplementary 1.5 for further explanation and comparison between the MH sampler and the AR6 original method.

Figure 5 also compares the emulation runs with the Chapter 4 assessment material, CMIP6 multi-model projections with observational constraints [37–39] and emulation runs configured with five different ECS-TCR pairs corresponding to the upper/lower bounds of *likely/very likely* ranges and best estimates (supplementary 1.4), which were combined with a 50–50 contribution as the AR6-assessed future warming projections.

The GSAT changes from the constrained CMIP6 and ECS-TCR mapped emulations generally agree well with each other, as well as with those from the four ensembles. However, it is noticeable that the ranges from the ECS-TCR mapped emulation are narrower in the long term than those from the constrained CMIP6 in low and very low emissions scenarios, labeled SSP1-2.6 and SSP1-1.9, respectively. The four ensembles lie between the two Chapter 4 elements with respect to range width. This difference implies that climate sensitivity alone is insufficient to comprehensively represent the uncertainty of GSAT changes. In fact, the ECS-TCR mapped emulation does not consider forcing uncertainty, which is important in lower-emissions scenarios, particularly for aerosols [40]. The emulation was run with each ensemble incorporating forcing perturbations, as in the constraining stage, which increases the GSAT range.

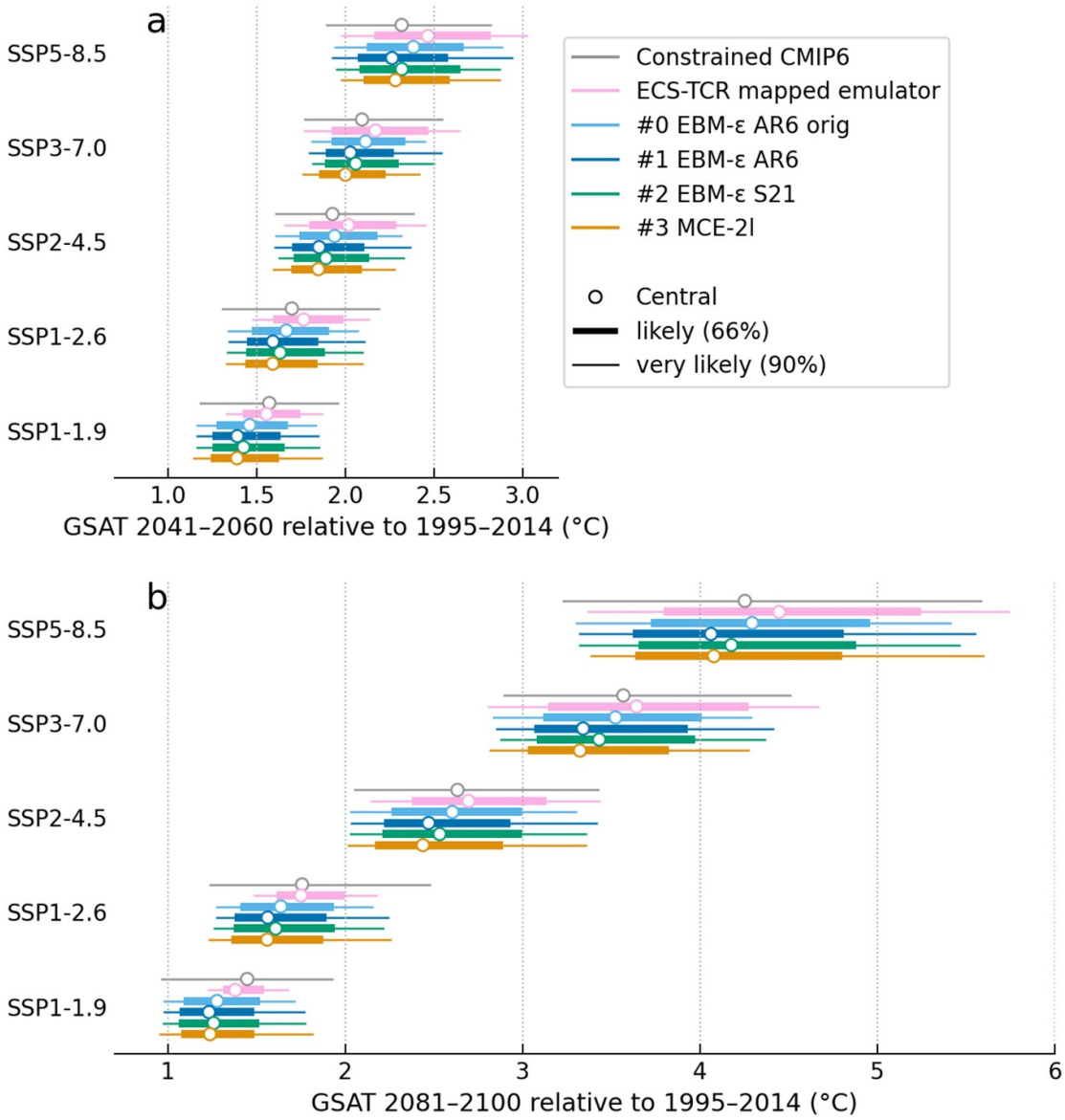

Figure 5. Ranges of global surface air temperature (GSAT) changes for mid-term (a, 2041–2060) and long term (b, 2081–2100) of five illustrative scenarios. The emulator ensembles are compared with evidence of the AR6-assessed future warming—constrained CMIP6 simulations, and two-layer model simulations configured with AR6-assessed ECS and TCR ranges, labeled as CMIP6 constrained and ECS & TCR mapping, respectively. GSAT changes are shown as changes relative to 1995–2014 plus 0.85 °C, the central estimate of the observed warming from 1850–1900 to 1995–2014 [41]. The scenarios are labeled as SSP*x*-*y*, where 'SSP*x*' refers to the Shared Socio-economic Pathways [42], and '*y*' refers to the approximate level of radiative forcing (in W m^{-2}).

Looking at the details, one notices that the ranges from the four ensembles are biased slightly lower than those from WGI Chapter 4, and that there are small but consistent differences between the four ensembles in terms of the magnitude of the central values. These differences are essentially a matter of constraint, and the ranges of the constrained parameters can be adjusted as necessary.

The emulation runs are characterized by forcing contributions, measured by forcing levels divided by λ, in conjunction with RWFs to changing forcing contributions, which depend on EBM parameters other than λ (supplementary figures 6 and 7). These diagnostic indicators explain the differences in GSAT between cases. Although the level of CO_2 forcing is generally smaller in Case 3 than in the others, its contribution to warming is rather close to each other owing to being divided by smaller λ. Warming generally occurs earlier in Case 3, which is associated with greater RWFs and also reduces the GSAT change differences between cases.

4. Discussion

Emulating ESMs often involves an 'out-of-sample' problem; that is, an emulator calibrated to a specific ESM with particular experiments does not always accurately represent other experiments with the same

ESM. A recent study on this problem provided several remedies, while highlighting the potential difficulties in improving accuracy [22]. The findings from the calibration stage in the present study, which are mainly related to the scaling of CO_2 forcing and ocean heat uptake, are in line with those of the previous study [22]. In fact, scaling CO_2 forcing to represent its super-logarithmic dependence on CO_2 concentration leads to a more general concept of forcing-specific efficacy [43], which is one of the remedies proposed in the previous study. The ocean heat uptake efficacy is one way to represent changing ΔN–ΔT slope under fixed forcing, termed as an 'effective' feedback parameter that evolves over time depending on changes in spatial patterns of forcing and response [44]. However, potential difficulty in representing the pattern effect to different forcing changes implies that single calibrated ε may not be suitable for general cases, leaving the standard EBM a feasible choice.

Aside from an estimation of an uncertain equilibrium state or a particular interest in historical changes in the effective climate feedback, the effect of including efficacy was observed for two distinctive time scales represented by a two-layer model. Calibrated time scales are generally longer with an EBM-ε than with a standard EBM, and longer time scales are attributed to a greater heat capacity of the model layers. In this regard, it should be noted that EBM-ε uses two different heat capacities—one for solving the two-layer energy balance equations, and the other for diagnosing ocean heat uptake, the latter of which is reduced by an ε factor (supplementary 1.1). Although this two-way use is a valid implementation of the equations, it may lead to physical ambiguity, particularly when emulating thermosteric sea level changes, as in AR6 WGI Chapter 9 [3]. One way to clarify the physical meaning and improve representation on a longer time scale would be to increase the number of model layers [23, 45, 46] with an extended period of calibration experiments [47]. However, that leaves issues regarding data processing for robust calibration.

Our future warming projections and those in the model intercomparison of RCMIP Phase 2 [48] both indicate that methodological differences in the calibration and sampling stages can be largely eliminated after the constraining stage. A similar temperature response corresponds to a similar ratio of forcing to feedback, which is not necessarily the same for each of forcing and feedback. This provides further insight into the causal relationship between the forcing and response. Currently, the concept of effective radiative forcing (ERF) is the basis of the forcing definition [49]. However, estimating the ERF from ESM experiments is not straightforward, and uncertainty is inevitable when separating forcing and feedback in accordance with the ERF definition, which requires special treatment for boundary conditions at the land and ocean surface [50]. The need for an advanced method to separate forcing from feedback was also pointed out in the previous study [22]. Introducing the concept of efficacy on the forcing side rather than the feedback side would be beneficial in this regard as a clue to fill the gaps between the standard ERF and ESM-specific forcing derived from calibration. In the present method, the scaling of non-CO_2 forcing agents was not considered, and the relationship between the AR6-assessed perturbed forcing levels and ESM-specific scaling was not examined. These issues should be addressed in future studies.

Although the above clarifications remain, the sampling stage should retain as much information of the base ESM ensemble as possible, assuming that the constraining stage can appropriately accept the members of the sampling results. It is rational to reflect the ESM ensemble in terms of its covariance structure in a multivariate parameter space, which also has the technical benefit of reducing member size. The constraining process critically affects probabilistic climate assessment as the final outcome, such as crossing times of 1.5 °C and 2 °C warming levels. The method used in this study is a tool that assists in the selection of appropriate constraining indicators in a standardized and flexible manner.

5. Conclusion

In the context of climate assessment, the most essential role of simple emulators, such as the two-layer energy balance model, is the synthesis of multiple lines of evidence, reflected in an ensemble of model parameters generated through calibration, sampling, and constraining processes. The present study intensively compared the AR6 method with three alternatives in terms of model formulation and calibrating procedures, fidelity to ESMs for forcing-response properties, and treatment of multivariate constraints.

Despite methodological differences, the constrained ensembles in these four cases showed similar climate projections in terms of central values and uncertainty ranges. The findings imply that constraining is the most critical process and is a high priority area in advancing the probabilistic approach. The MH sampler used in the alternative cases would be a promising option, which enables setting multivariate constraints in a standardized and flexible manner.

Being properly adjusted in the final constraining also implies that there is more flexibility in the prior processes. In sampling for unconstrained ensembles, there is a tradeoff between reducing bias and retaining consistency. While the original AR6 method partially introduces bias-free independent parameter series, the alternative inherits the statistics of the CMIP6 calibrated ensemble for all parameters. Since bias reduction can be left to constraining, there would be an

advantage to the latter consistent alternative. The heat uptake efficacy term in the two-layer model improves the accuracy of emulation under certain conditions but may not work robustly under all conditions. Its practical use may be limited to specific purposes, such as the evaluation of ECS or transient effective climate feedback. Model improvement with a kind of efficacy can be approached from the forcing side of the two-layer model. Scaling CO_2 forcing and the additional super-logarithmic scheme, either or both used in the alternative cases, were beneficial for accuracy and would be extended to non-CO_2 forcing.

Through revisiting the probabilistic approach with the AR6 two-layer model, the present study has identified improvements in its design and implementation. The findings can be reflected in better synthesis of multiple evidence in the next assessment cycle, leading to a more robust mitigation scenario assessment.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.12845227 [51].

Acknowledgments

This work was supported by the MEXT (Ministry of Education, Culture, Sports, Science and Technology, Japan) Program for Advanced Studies of Climate Change Projection (SENTAN), Grant Number JPMXD0722681344. The CMIP6 outputs used in this study were obtained from the Earth System Grid Federation (supplementary table 1).

ORCID iDs

Junichi Tsutsui https://orcid.org/0000-0003-1112-4335
Chris Smith https://orcid.org/0000-0003-0599-4633

References

[1] Forster P *et al* 2021 The Earth's energy budget, climate feedbacks, and climate sensitivity *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* ed V Masson-Delmotte *et al* (Cambridge University Press) pp 923–1054
[2] Lee J Y *et al* 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* ed V Masson-Delmotte *et al* (Cambridge University Press) pp 553–672
[3] Fox-Kemper B *et al* 2021 Ocean, cryosphere and sea level change *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* ed V Masson-Delmotte *et al* (Cambridge University Press) pp 1211–362
[4] Held I M, Winton M, Takahashi K, Delworth T, Zeng F and Vallis G K 2010 *J. Clim.* **23** 2418–27
[5] Geoffroy O, Saint-Martin D, Bellon G, Voldoire A, Olivié D J L and Tytéca S 2013 *J. Clim.* **26** 1859–76
[6] Winton M, Takahashi K and Held I M 2010 *J. Clim.* **23** 2333–44
[7] Stevens B, Sherwood S C, Bony S and Webb M J 2016 *Earth's Future* **4** 512–22
[8] Geoffroy O, Saint-Martin D, Olivié D J L, Voldoire A, Bellon G and Tytéca S 2013 *J. Clim.* **26** 1841–57
[9] Nicholls R J *et al* 2020 *Geosci. Model Dev.* **13** 5175–90
[10] Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 *Geosci. Model Dev.* **9** 1937–58
[11] Meinshausen M, Meinshausen N, Hare W, Raper S C B, Frieler K, Knutti R, Frame D J and Allen M R 2009 *Nature* **458** 1158–62
[12] Rogelj J, Meinshausen M and Knutti R 2012 *Nat. Clim. Change* **2** 248–53
[13] Schaeffer M, Gohar L, Kriegler E, Lowe J, Riahi K and van Vuuren D 2015 *Technol. Forecast Soc.* **90** 257–68
[14] Smith C, Nicholls Z R J, Armour K, Collins W, Forster P, Meinshausen M, Palmer M D and Watanabe M 2021 The Earth's energy budget, climate feedbacks, and climate sensitivity. Supplementary material *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* ed V Masson-Delmotte *et al* (Cambridge University Press) pp 923–1054
[15] Smith C, Cummins D P, Fredriksen H B, Nicholls Z, Meinshausen M, Allen M, Jenkins S, Leach N, Mathison C and Partanen A I 2024 *Geosci. Model Dev.* **17** 8569–92
[16] Clarke L *et al* 2014 Assessing transformation pathways *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* ed O Edenhofer *et al* (Cambridge University Press) pp 413–510
[17] Rogelj J *et al* 2018 Mitigation pathways compatible with 1.5 °C in the context of sustainable development *Global Warming of 1.5 °C an IPCC Special Report on the Impacts of Global Warming of 1.5 °C above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty* ed V Masson-Delmotte *et al* (Cambridge University Press) pp 93–174
[18] Riahi K *et al* 2022 Mitigation pathways compatible with long-term goals *Climate Change 2022: Mitigation of Climate Change Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* ed P R Shukla *et al* (Cambridge University Press) pp 295–408
[19] Smith C J, Forster P M, Allen M, Leach N, Millar R J, Passerello G A and Regayre L A 2018 *Geosci. Model Dev.* **11** 2273–97
[20] Kikstra J S *et al* 2022 *Geosci. Model Dev.* **15** 9075–109
[21] O'Neill B C *et al* 2016 *Geosci. Model Dev.* **9** 3461–82
[22] Jackson L S, Maycock A C, Andrews T, Fredriksen H B, Smith C J and Forster P M 2022 *Geophys. Res. Lett.* **49** e2022GL098808
[23] Cummins D P, Stephenson D B and Stott P A 2020 *Adv. Stat. Clim. Meteorol. Oceanogr.* **6** 91–102
[24] Tierney L 1994 *Ann. Stat.* **22** 1701–28
[25] Smith C J *et al* 2021 *J. Geophys. Res.* **126** e2020JD033622
[26] Tsutsui J 2020 *Geophys. Res. Lett.* **47** e2019GL085844
[27] Tsutsui J 2022 *Geosci. Model Dev.* **15** 951–70
[28] Andrews T, Gregory J M, Webb M J and Taylor K E 2012 *Geophys. Res. Lett.* **39** L09712
[29] Boucher O *et al* 2020 *J. Adv. Modeling Earth Syst.* **12** e2019MS002010

[30] Meraner K, Mauritsen T and Voigt A 2013 *Geophys. Res. Lett.* **40** 5944–8
[31] Meinshausen M *et al* 2020 *Geosci. Model Dev.* **13** 3571–605
[32] Rugenstein M A A, Gregory J M, Schaller N, Sedláček J and Knutti R 2016 *J. Clim.* **29** 5643–59
[33] Fredriksen H B, Smith C J, Modak A and Rugenstein M 2023 *Geophys. Res. Lett.* **50** e2023GL102916
[34] Gregory J M, Ingram W J, Palmer M A, Jones G S, Stott P A, Thorpe R B, Lowe J A, Johns T C and Williams K D 2004 *Geophys. Res. Lett.* **31** L03205
[35] Knutti R, Rugenstein M A A and Hegerl G C 2017 *Nat. Geosci.* **10** 727–36
[36] Pfister P L and Stocker T F 2018 *Environ. Res. Lett.* **13** 124024
[37] Liang Y, Gillett N P and Monahan A H 2020 *Geophys. Res. Lett.* **47** e2019GL086757
[38] Tokarska K B, Stolpe M B, Sippel S, Fischer E M, Smith C J, Lehner F and Knutti R 2020 *Sci. Adv.* **6** eaaz9549
[39] Ribes A, Qasmi S and Gillett N P 2021 *Sci. Adv.* **7** eabc0671
[40] Watson-Parris D and Smith C J 2022 *Nat. Clim. Change* **12** 1111–3
[41] Gulev S K *et al* 2021 Changing state of the climate system *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* ed V Masson-Delmotte *et al* (Cambridge University Press) pp 287–422
[42] Riahi K *et al* 2017 *Glob. Environ. Change* **42** 153–68
[43] Hansen J *et al* 2005 *J. Geophys. Res.* **110** D18104
[44] Goodwin P 2018 *Earth's Future* **6** 1336–48
[45] Tsutsui J 2017 *Clim. Change* **140** 287–305
[46] Fredriksen H B and Rypdal M 2017 *J. Clim.* **30** 7157–68
[47] Rugenstein M *et al* 2019 *Bull. Am. Meteorol. Soc.* **100** 2551–70
[48] Nicholls Z *et al* 2021 *Earth's Future* **9** e2020EF001900
[49] Forster P M, Richardson T, Maycock A C, Smith C J, Samset B H, Myhre G, Andrews T, Pincus R and Schulz M 2016 *J. Geophys. Res.* **121** 12,460–75
[50] Smith C J *et al* 2020 *Atmos. Chem. Phys.* **20** 9591–618
[51] Tsutsui J 2024 Revisiting the two-layer energy balance model used in IPCC AR6, v1.0.0 *Zenodo* https://doi.org/10.5281/zenodo.12845227