ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Long run emulator calibration increases warming and sea-level rise projections

To cite this article: Christopher D Wells *et al* 2026 *Environ. Res. Lett.* **21** 034008

View the article online for updates and enhancements.

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
14 October 2025

REVISED
20 December 2025

ACCEPTED FOR PUBLICATION
14 January 2026

PUBLISHED
30 January 2026

LETTER

Long run emulator calibration increases warming and sea-level rise projections

Christopher D Wells[1,*], Donald P Cummins[1,2], Haozhe He[3] and Chris Smith[4]

[1] School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom
[2] Centre for Environmental Modelling And Computation (CEMAC), University of Leeds, Leeds LS2 9JT, United Kingdom
[3] High Meadows Environmental Institute, Princeton University, Princeton, NJ, United States of America
[4] Energy, Climate and Environment Program, International Institute for Applied Systems Analysis (IIASA), 2361 Laxenburg, Austria
[*] Author to whom any correspondence should be addressed.

E-mail: c.d.wells@leeds.ac.uk

Keywords: simple climate modelling, climate emulation, emulator calibration, overshoot scenarios

Supplementary material for this article is available online

Abstract

Owing to their short runtime compared to Earth system models (ESMs), as well as the difficulty for the latest ESMs from the Coupled Model Intercomparison Project Phase 6 (CMIP6) to reproduce historical warming and the so-called 'hot model problem', constrained reduced-complexity climate models ('emulators') are increasingly used to produce global warming projections from emissions scenarios. Emulators are often calibrated on idealised abrupt CO_2 quadrupling experiments from CMIP6, particularly the global surface temperature response over time to an imposed radiative forcing. Such CMIP6 experiments tend to be run for 150 years, which is not sufficient to reveal the full equilibrium response to an imposed climate forcing. Here we show that, when longer experiments are available for emulator calibration, the long-term climate warming projections increase, particularly for 2100, by up to 0.70 (0.42–0.93, 25th to 75th percentile) °C in the median under a high emissions scenario; peak global warming in a high overshoot scenario is higher by 0.24 °C (0.14–0.31 °C). Corresponding long-term thermosteric sea level rise (SLR) is consequently higher, by 0.45 (0.22–0.52, 25th to 75th percentile) m in 2500. This result, consistent across calibrations from 17 ESMs, has implications for climate change mitigation strategies, as it is likely that even more stringent emissions reductions would be required to limit long-term warming and SLR than previously thought.

1. Introduction

Reduced-complexity climate models and climate model emulators have found utility in fields such as economics (Nordhaus 1991) and integrated assessment modelling (Kikstra *et al* 2022, Riahi *et al* 2022) for decades, but are increasingly used to make inferences about physical climate change (Forster *et al* 2021), particularly in response to future emissions projections (Lee *et al* 2021). Emulators are typically calibrated to the large-scale behaviour of more complex climate models (Meinshausen *et al* 2011, Dorheim *et al* 2024, Sandstad *et al* 2024), and often constrained against observations such as global mean surface temperature (GMST) (Smith *et al* 2024, Tsutsui and Smith 2024), to ensure that the climate projections they produce are plausible. Within the Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC), it was identified that many of the GMST outputs from the ensemble of ESMs contributing to CMIP6 did not correspond well with historical observations and were warmer than expected in the future, in many cases due to high rates of recent warming (Flynn and Mauritsen 2020, Hausfather *et al* 2020, Lee *et al* 2021, Smith and Forster 2021). The ensemble of CMIP6 models tended to show higher equilibrium climate sensitivity (ECS) than those in the previous CMIP5 ensemble (Zelinka *et al* 2020), with several above the 'very likely' upper bound of 5 °C assessed by the IPCC AR6 (Forster *et al* 2021). This led to warmer projections than expected from future emissions scenarios

(Tebaldi *et al* 2021), and caused the IPCC to base their future warming assessments on lines of evidence that, for the first time, did not include unadjusted CMIP model results (Lee *et al* 2021). Climate emulators such as the two-layer energy balance model (EBM; Held *et al* 2010, Winton *et al* 2010, Geoffroy *et al* 2013a) that model the response of the near-surface and deep ocean warming to a radiative forcing over time, became an important line of evidence to assess future warming and sea-level rise (Fox-Kemper *et al* 2021, Lee *et al* 2021, Kopp *et al* 2023).

The EBM is usually calibrated to the response of the *abrupt-4xCO₂* experiment (Cummins *et al* 2020, Geoffroy *et al* 2013b, Smith *et al* 2021b, 2024, Tsutsui and Smith 2024), which has been a staple diagnostic experiment across CMIP generations since CMIP5 (Taylor *et al* 2012), with variants going back to the 1979 Charney assessment of ECS (Charney *et al* 1979). In this experiment, the atmospheric concentration of CO_2 is abruptly quadrupled above pre-industrial levels and the model run for typically 150 years (Eyring *et al* 2016). This allows for estimates of the climate feedback, effective radiative forcing (due to CO_2), and ECS to be obtained from ESMs (Gregory *et al* 2004). Noting that there is often an increase in climate sensitivity over time in ESMs due to a forced 'pattern effect' of sea surface temperature distributions (Andrews *et al* 2015), introducing a parameter representing the efficacy of deep-ocean heat uptake allows for different effective climate feedback strengths in the short and longer term (Geoffroy *et al* 2013a), and more accurately simulates the warming profiles of most ESMs (Dai *et al* 2020).

However, 150 years is not long enough for the climate state to reach an equilibrium (Rugenstein *et al* 2019, 2020), which may take thousands of years and is governed by the characteristic response timescale of the deep ocean (Li and Jarvis 2009, Yang and Zhu 2011). Therefore, in 150-year *abrupt-4xCO₂* experiments, ECS is typically estimated by a regression of the annual mean GMST against top-of-atmosphere radiative imbalance (N) and taking the intercept at $N = 0$ (and since ECS is defined as the warming from a doubling of CO_2 rather than a quadrupling, this value is divided by two). This linear regression-derived value is known as the effective climate sensitivity (EffCS) (Andrews *et al* 2018). Owing to the forced pattern effect, climate sensitivity typically increases over time, and performing the same regression calculation for longer model runs typically yields higher EffCS values (Rugenstein *et al* 2020).

Although 150 years is not long enough to resolve the long-term response of ESMs, due to data availability the climate emulation community often uses EBMs calibrated on the 150-year *abrupt-4xCO₂* experiment and apply these results to scenario forcings spanning the historical and future period which is much longer than 150 years (for example, 1750 or 1850 to 2300 or 2500). Therefore, such emulators may perform sufficiently well over the historical period (less than two calibration timeframes for 1750–2025), but struggle to resolve time horizons much longer when considering both GMST (Jackson *et al* 2022) and sea-level rise (Malagón-Santos *et al* 2025). This can be concluded by noting that the characteristic timescales of the slowest response mode of three-layer model calibrations typically increase with increasing length of *abrupt-4xCO₂* training data (supplementary figure S4).

A climate modelling protocol, longRunMIP (Rugenstein *et al* 2019), calls for *abrupt-4xCO₂* experiments to be run for longer time periods, a suggestion also endorsed by Fredriksen *et al* (2025). Dai *et al* (2020) showed that for three participating models in longRunMIP, using 150 years to calibrate a two-layer EBM resulted in substantially cooler long-term projections for *abrupt-4xCO₂* experiments than using 1000 years, which was much closer to the results obtained using all available years in each model (around 5000). In this paper we calibrate a three-layer EBM to these longer timeframe experiments, collated from longRunMIP and other initiatives, and show that future warming and sea-level rise in climate scenarios tends to be higher in longer three-layer calibrations relative to 150 year calibrations. We also compare the 150 year response between two- and three-layer models, finding that projections are not systematically affected by increasing from two to three layers. We discuss the implications for 21st century climate projections and beyond using these models.

2. Method

2.1. Two- and three-layer EBMs

A widely used climate model emulator is the k-layer EBM that projects the change in top-of-atmosphere energy imbalance N and temperature anomalies in each ocean layer $T_1, \ldots T_k$ as a response to an imposed radiative forcing F. The formulation here largely follows the stochastic k-layer EBM described by Cummins *et al* (2020), from which we drop the stochastic components. The GMST anomaly is taken to be equivalent to the temperature of the uppermost layer (T_1) given the substantially larger heat capacity of the near-surface ocean layer compared to the atmosphere and land surface (von Schuckmann *et al* 2020). The so-called deep ocean heat uptake efficacy factor ε is implemented to model the delayed response of the deep ocean warming that increases climate sensitivity over time (Geoffroy *et al* 2013a, Cummins *et al* 2020).

The two-layer EBM can be written as

$$C_1 \frac{\mathrm{d}T_1}{\mathrm{d}t} = F - \kappa_1 T_1 - \varepsilon \kappa_2 (T_1 - T_2)$$

$$C_2\frac{\mathrm{d}T_2}{\mathrm{d}t} = \kappa_2\left(T_1 - T_2\right)$$

and the three-layer EBM as

$$C_1\frac{\mathrm{d}T_1}{\mathrm{d}t} = F - \kappa_1 T_1 - \kappa_2\left(T_1 - T_2\right)$$

$$C_2\frac{\mathrm{d}T_2}{\mathrm{d}t} = \kappa_2\left(T_1 - T_2\right) - \varepsilon\kappa_3\left(T_2 - T_3\right)$$

$$C_3\frac{\mathrm{d}T_3}{\mathrm{d}t} = \kappa_3\left(T_2 - T_3\right).$$

In both cases C_i, κ_i $(i = 1, \ldots, k)$ represent the heat capacity and heat transfer coefficients of each ocean layer respectively. κ_1 can also be expressed as $-\lambda$, where λ is the climate feedback parameter.

The relationship between top-of-atmosphere energy imbalance N and GMST T_1 is governed by

$$N = F - \kappa_1 T_1 + (1-\varepsilon)\kappa_k\left(T_{k-1} - T_k\right).$$

The calibration of the k-layer EBM is performed using non-linear least squares with L2 regularisation on characteristic timescales. For a k-layer model there are $2k + 2$ free parameters to fit ($\kappa_1, \ldots, \kappa_k, C_1, \ldots, C_k, F_{4\times\mathrm{CO2}}, \varepsilon$). In the calibration data from the *abrupt-4xCO$_2$* experiments, the model has visibility of only the N and T_1 state variables. The assumption of a constant (but *a priori* unknown) forcing of $F = F_{4\times\mathrm{CO2}}$ over all times $t > 0$ is imposed as a boundary condition.

2.2. Impulse response form and climate sensitivity metrics

The k-layer EBM formulation is mathematically equivalent to a k_{th}-order impulse response model (IRM) using k thermal boxes characterised by their response coefficients q_{ij} and characteristic timescales τ_j (Fredriksen and Rypdal 2017, Tsutsui 2017, Leach *et al* 2021). The partial contributions $S_{ij}(t)$ to the total temperature change $T_i(t)$ in physical layer i and for thermal box j can be written in the form

$$\frac{\mathrm{d}S_{ij}(t)}{\mathrm{d}t} = \frac{q_{ij}F(t) - S_{ij}(t)}{\tau_j}$$

and

$$T_i(t) = \sum_{j=1}^{k} S_{ij}(t).$$

From here on we focus on the near-surface layer $i = 1$, and drop the subscript i.

The characteristic timescales τ_j can be interpreted as the response timescales of k independent modes, where the longest timescale τ_k is the projection of the deep ocean mode on the surface temperature. The response coefficients q_j define the strength of the contribution of each mode onto the near-surface temperature T_1.

From the impulse response formulation, the conventional climate sensitivity metrics of ECS and transient climate response (TCR) can be derived. The ECS is defined as the long-term equilibrium temperature anomaly following a doubling of CO$_2$. The ECS we calculate is the true equilibrium value that would be obtained when the EBM is run for a sufficiently long time, rather than a regression-based EffCS (Gregory *et al* 2004, Andrews *et al* 2018) typical of ESM-derived estimates. ECS is estimated from the EBM parameters as

$$\mathrm{ECS} = \frac{F_{4\times\mathrm{CO}_2}}{2\kappa_1}$$

and equivalently from the IRM parameters as

$$\mathrm{ECS} = \frac{F_{4\times\mathrm{CO}_2}}{2}\sum_{j=1}^{k} q_j,$$

where the division by two scales the forcing from a quadrupling to a doubling of CO$_2$, under the assumption of a perfectly logarithmic relationship between CO$_2$ concentration and radiative forcing.

The TCR is not a true TCR as estimated from 70 years of a compound 1% per year CO$_2$ increase experiment as is typically performed in ESMs (termed *1pctCO$_2$*; Eyring *et al* (2016)), but can be approximated from the IRM parameters (Jiménez-de-la-Cuesta and Mauritsen 2019) as

$$\mathrm{TCR} = \frac{F_{4\times\mathrm{CO}_2}}{2}\sum_{j=1}^{k} q_j\left(1 - \frac{\tau_j}{D}\left(1 - \mathrm{e}^{-\frac{D}{\tau_j}}\right)\right)$$

where $D = \log(2)/\log(1.01) \approx 69.7$ yr, the time taken to reach a doubling of CO$_2$ following a rate of 1% per year compound increase. The approximate and 'true' TCR values derived from a *1pctCO$_2$* experiment using an EBM-based simple climate model are very similar, as demonstrated in Smith *et al* (2024).

2.3. Longrunmip data

We use 17 ESMs that contributed *abrupt-4xCO$_2$* runs to longRunMIP and CMIP6 (Eyring *et al* 2016, Rugenstein *et al* 2019), providing between 500 and 5900 years of data (see supplementary table S1 for models and experiment lengths), using the CMIP variable names *tas* (global mean surface air temperature) and *rtmt* (net radiation imbalance at the top of atmosphere). *tas* corresponds to T_1 and *rtmt* to N in the EBM framework. Following recommendations in the longRunMIP protocol, neither *tas* nor *rtmt* is de-drifted, as the models are verified to be in approximate radiative and thermal equilibrium at the start of the experiment, and many models did not run a parallel pre-industrial control run for as long as the *abrupt-4xCO$_2$* experiment that would be necessary to properly apply any dedrifting method (Zehrung *et al* 2025).

We calibrate a three-layer EBM to the full length of each model's available *abrupt-4xCO$_2$* run, plus the first 150, 300, 500, 900, 1000, 1800, 2000, 3000, 4000, and 5000 years of output for models running at least this number of years. In addition, we calibrate a two-layer EBM for the first 150 years of data, for comparison with the three-layer response over the same period. The first 150 years is in line with the CMIP *abrupt-4xCO$_2$* time horizon, and we use successively longer periods to determine the additional information obtained from extra data. All models ran for at least 500 years, and all but two ran for at least 999 years, which is why this study prioritises the 900 year results. Calibrated parameters for each model, calibration length and number of layers are provided in the extended data table 1, as well as their impulse-response function forms which are mathematically equivalent (Leach *et al* 2021, Nicholls and Lewis 2021). In general, we observe that the characteristic timescale of the deep ocean (the longest timescale in the three-time constant impulse response form) increases with the number of available calibration years. This is expected, as the true equilibrium time of the deep ocean is estimated to be many centuries (Li and Jarvis 2009), with little response over the first 150 years of the response (Tsutsui 2017), and shorter calibration periods such as 150 years are unable to resolve timescales several times larger than the calibration data.

2.4. Scenario forcing data

To estimate the climate response to historical and future emissions in each ESM, we drive the EBM calibrations with the effective radiative forcing time series provided by the IPCC AR6 WG1 (Smith *et al* 2021a, 2021b) spanning the historical period (1750–2019) and eight shared socioeconomic pathways coupled with representative concentration pathways for 2020–2500. For the analysis in figure 3, we focus mainly on ssp119 (1.5 °C aligned), ssp245 (approximately current policies), ssp534-over (high overshoot and reversal) and ssp585 (very high emissions). We also produce projections from ssp126, ssp370, ssp434 and ssp460, which, in combination from the other four SSPs, we use for analysis of global warming level (GWL) crossing times in figure 4. We disregard the fact that the IPCC forcing time series does not necessarily correspond to what each model would 'see', given the differing responses of ESMs to forcings arising from their radiative transfer parameterisations and meteorology (Smith *et al* 2020), and as evaluating each model's scenario forcing response is not part of the longRunMIP protocol. This also renders meaningful comparisons of EBM to ESM data impossible for these scenarios. We do not constrain the EBM outputs to match historical climate observations, but use GMST and SLR projections relative to 2020 to evaluate the differences in projected future warming between different calibration lengths.

2.5. GWL crossing times

The crossing time of different GWLs under each scenario is evaluated by defining the 2015–2024 GMST anomaly as 1.24 °C relative to pre-industrial following the Indicators of Global Climate Change (Forster *et al* 2025), and evaluating the central year of a 20 year mean that crosses each given GWL following the logic of IPCC AR6 (Lee *et al* 2021). This is done separately for projections of all five Tier 1 and three of the four Tier 2 SSP scenarios (Meinshausen *et al* 2020; excluding ssp370-lowNTCF): ssp119, ssp126, ssp245, ssp370, ssp585, ssp434, ssp460, and ssp534-over.

2.6. Calculation of thermosteric sea level rise (SLR)

We convert ocean heat uptake of energy to the corresponding level of thermosteric SLR using the conversion factor of 0.0975 m YJ^{-1} found by Ramme *et al* (2025).

3. Results

Firstly, we find consistently higher ECS values when training the three-layer EBM on *abrupt-4xCO$_2$* experiments of increasing length (figure 1; see Method for our calculation of ECS). Overall, 16 of 17 ESMs show higher ECS values when training on all available data (between 500 and 5900 years; see supplementary table S1 and extended data table 1), with 14 of the 15 ESMs with 900+ years of data featuring higher ECS when trained on 900 years as compared to 150 years. In both cases, HadGEM2 is the outlier showing a reduced ECS with longer training length. HadGEM2's *abrupt-4xCO$_2$* emulation using shorter calibrations is poor (substantially too warm; supplementary figure S1), suggesting the full response is inadequately captured in the first 150 years. To facilitate comparison between consistent sets of models when analysing data of 900 years or more, the two models with runs shorter than 900 years, NorESM2-LM and GFDL-ESM4, are dropped from the multi-model analysis.

We can run the calibrated EBMs with radiative forcing time series from emissions scenarios to make future climate projections. Figure 2 shows the difference in historical and future GMST trajectories (20 year rolling average) across 8 scenarios

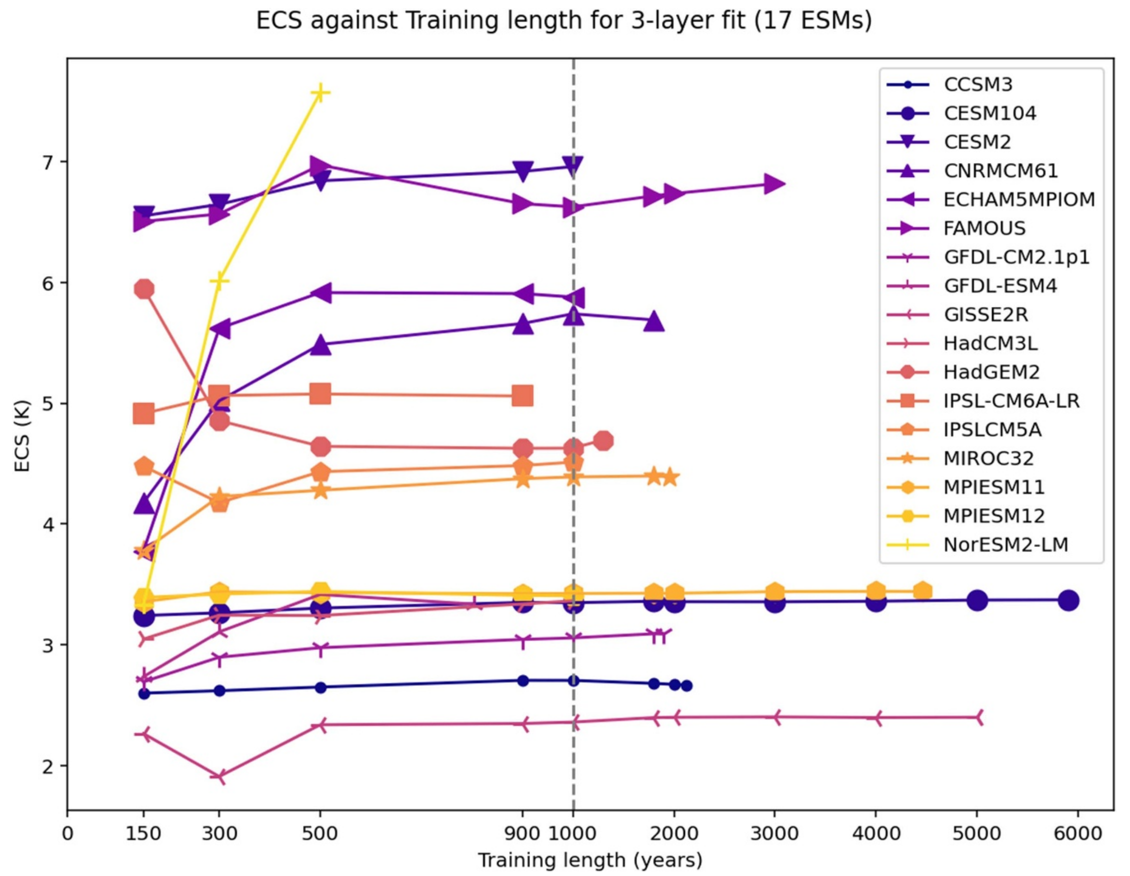

Figure 1. Calculated ECS values determined from a three-layer energy balance model (EBM) for 17 ESMs as a function of the length of the *abrupt-4xCO$_2$* data used to train each EBM. Note that the horizontal scale changes at 1000 years (vertical grey dashed line) to enable visualisation of the full data.

between different calibration lengths and number of layers in the EBM (see supplementary figure S2 for the actual GMST anomaly responses). Differences in the historical period are minimal, with calibrations generally varying by substantially less than a few tenths of a degree throughout. However, this historical behaviour contrasts markedly with the future projections, which differ markedly between calibration approaches. While increasing the EBM from two to three layers (solid lines) has little systematic effect, increasing the three-layer calibration length from 150 to 300 (dashed lines) or 900 (dotted lines) years increases temperature projections in the medium-term, particularly around 2100, and generally (though with some exceptions) also in the long term. Across the wide range of scenarios, we consistently find higher mid-term warming upon extending the calibration (dotted, dashed lines).

In figure 3, we compare different future warming and thermosteric SLR projections relative to 2020 for four climate scenarios. We find little systematic effect on warming across periods when increasing from two to three EBM layers in the 150 year fit (solid bars, figure 3), and slightly higher near-term warming when extending the three-layer EBM training data from 150 to 300 years (hatched bars, figure 3). In contrast, we find large increases in warming when extending the training data from 300 to 900 years in the three-layer EBM (cross-hatched bars, figure 3). This suggests that 300 years of data is insufficient to capture the full response to the *abrupt-4xCO$_2$* forcing. Projections of thermosteric SLR are slightly higher on multi-century timescales when extending from two to three layers or from 150 to 300 years of training data (figure 3), due to the higher accumulation of ocean heat during periods of positive temperature anomalies. Substantial rises in SLR projections occur when shifting to the 900 year three-layer fit, consistent with the greater GMST anomalies.

The warmer projections when using longer training data affect key properties of future temperature trajectories (figure 4). When we base climate projections under eight SSP scenarios to the same recent warming estimate (see Methods), long-term

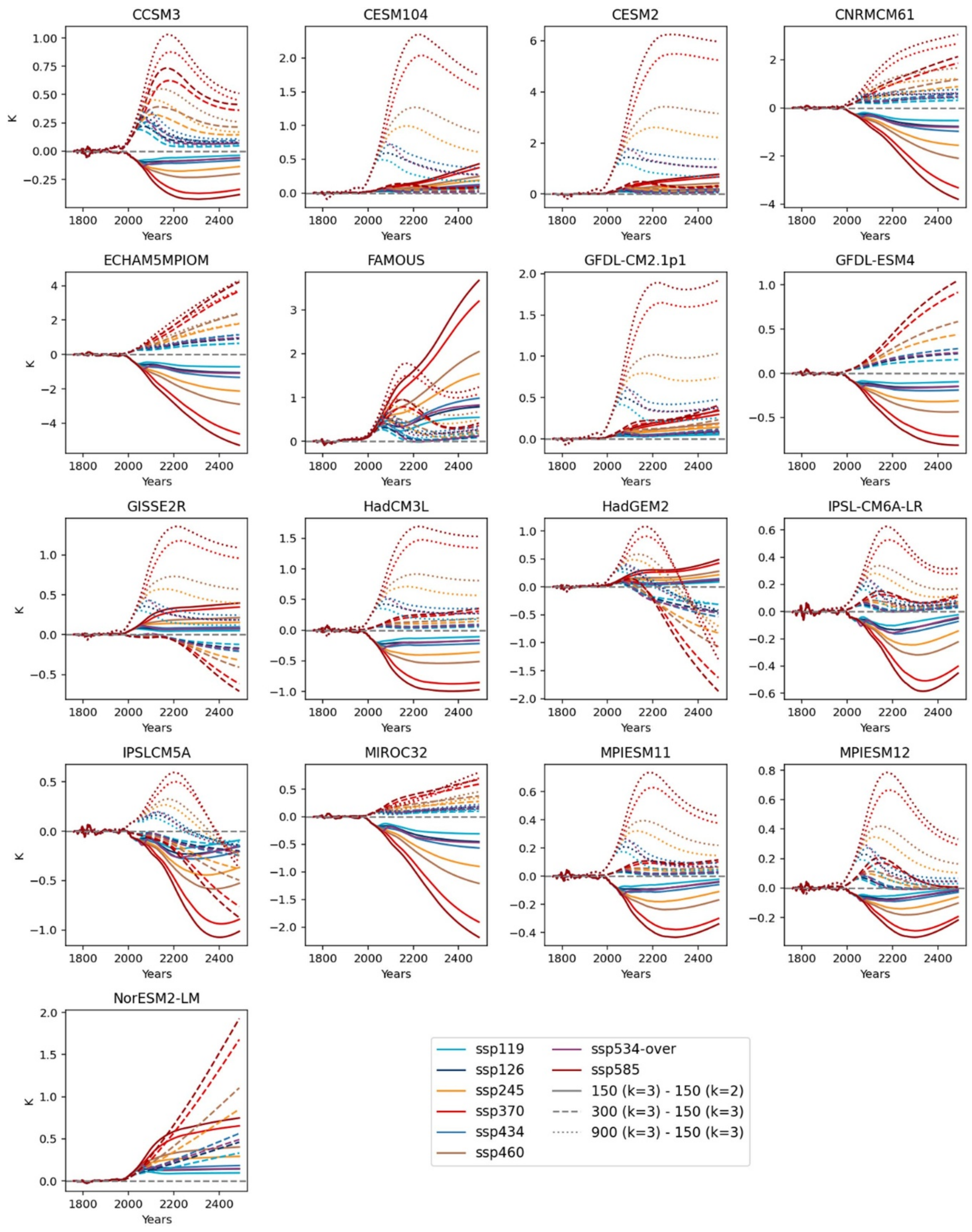

Figure 2. Effect of EBM training length and number of layers of time series of GMST responses to radiative forcing for each SSP (using the same forcing for each case; see Method) for each ESM. Data is the 20 year rolling average from 1750 to 2500. Note that the 900 year comparisons are not available for GFDL-ESM4 and NorESM2-LM (Methods).

GWLs are reached earlier. Crossing times for higher GWLs consistently occur several years earlier when using 900 years of calibration data in the three-layer model compared to 150 years; 1.5 °C and 2 °C crossing times are up to 2 years earlier in the median response, though some ESM calibrations see shifts of several years (figure 4(a)). We find that magnitudes of peak warming in overshoot (ssp119 and ssp534-over) scenarios are more substantially affected, with the very low-emissions ssp119 scenario 0.07 °C (25th to 75th percentile: 0.06 °C–0.09 °C) warmer when increasing from 150 to 900 years in a three-layer EBM, and ssp534-over warmer by 0.24 °C (0.15 °C–0.31 °C) (figure 4(b)). Increasing from 150 to 300 or 900 years of data in the three-layer model results in increased peak warming under both overshoot scenarios in almost all ESMs.

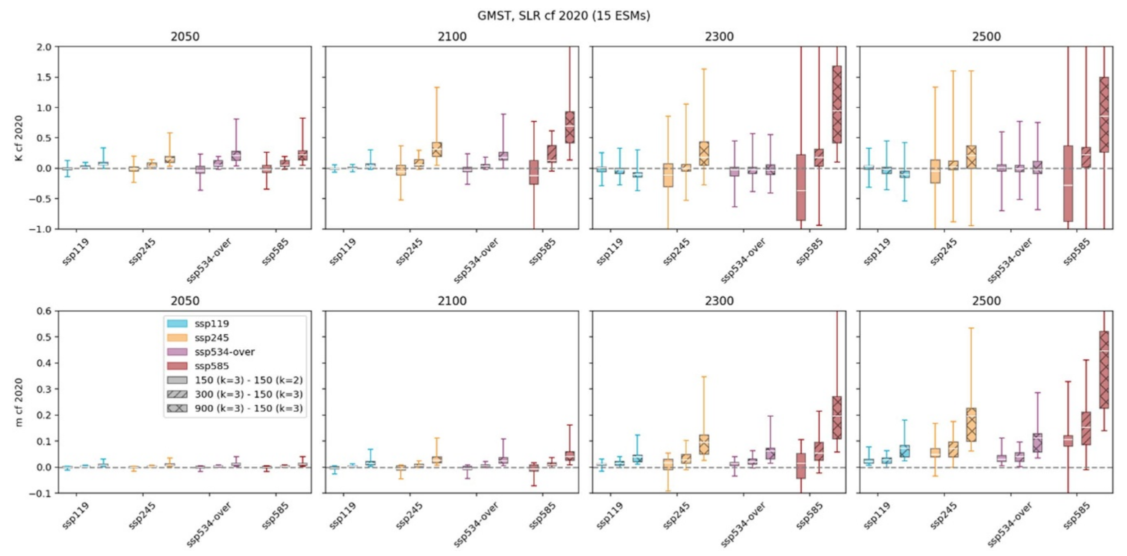

Figure 3. Changes in GMST (top) and SLR (bottom) in 2050, 2100, 2300 and 2500 relative to 2020, in the ssp119, ssp245, ssp534-over and ssp585 scenarios upon utilising different lengths of training data and numbers of layers in the EBM. Results are shown comparing different EBM fits. Box-and-whisker plots show the median and interquartile range, and the minimum and maximum values, across the 15 ESMs which ran 900-plus years of *abrupt-4xCO*$_2$ training data.

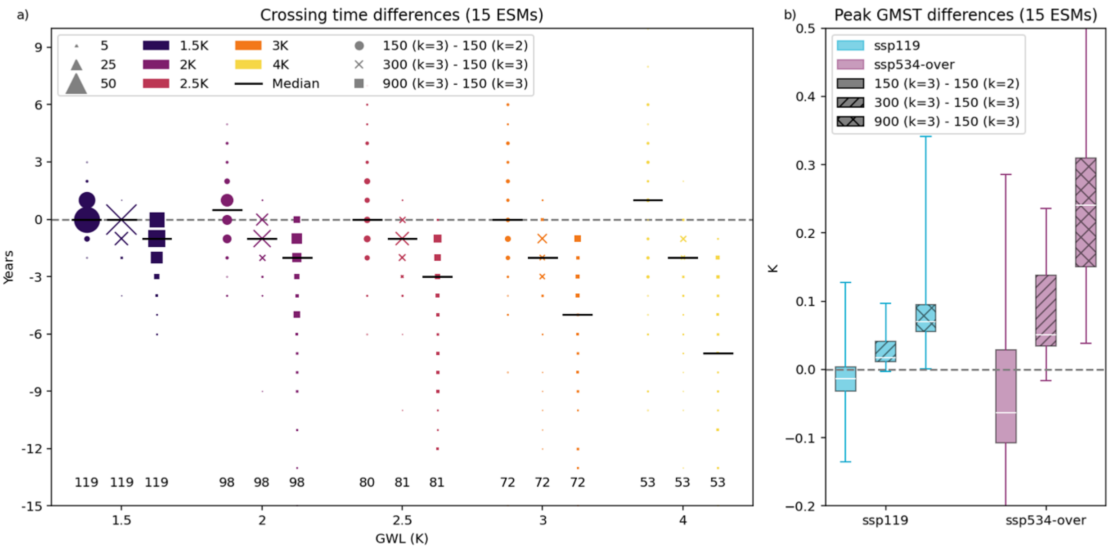

Figure 4. Effect of different training lengths and EBM layers on the crossing time of various global warming levels (a) and on peak warming (b). Crossing time differences are aggregated across eight SSP scenarios (see Method). Since the driving radiative forcing time series are annually resolved and the EBM is run at an annual timestep, crossing time differences are integer years, and are therefore shown as discrete data, with the median value shown in black. Marker sizes reflect the number of calibration-scenario combinations exhibiting each change in crossing time. The number of crossing occurrences (ESM-SSP combinations) in which each GWL is reached for each comparison is shown underneath. Variations in the level of peak warming in two overshoot scenarios are calculated, with box-and-whisker plots showing the median and interquartile range and the minimum and maximum values.

4. Discussion and conclusions

Our main finding is that calibrating climate emulators on 150 years of model data is insufficient to capture the long-term response to forcing, with more stable climate parameters when using more data. In particular, we find higher values of climate sensitivity and deep-ocean response timescale when calibrating an EBM to longer time series. Using the 150 year period leads to low-biased estimates of GMST, particularly during the 21st century, and of sea-level rise during the longer term. We also find that the CMIP7 recommendation of running *abrupt-4xCO*$_2$ experiments for 300 years (Dunne *et al* 2025) does little to resolve this problem, with simulations calibrated on 900 years of data showcasing higher temperatures and SLR projections.

We therefore recommend that, ideally, *abrupt-4xCO$_2$* experiments be performed for 1000 years to more appropriately resolve the long-term thermal responses of the climate system and better calibrate climate model emulators. Additionally, the instability of the calibrated response even on timescales of thousands of years (with substantial changes in ECS and projections when increasing past 1000 years; see figure 1 and supplementary figures S1, S4) motivates the simulation of multi-millennial runs, where computing power permits.

We find in addition that adding a third layer tends to improve the performance of the EBM on both short and long timescales compared to the two-layer model, as the extra layer in the model can in principle resolve both short-term and long-term dynamics. The three-layer fit results in lower root-mean-square errors taken across the full period in 63% of cases, and in 67% of cases when considering the final 100 years of the simulation (figure S3). However, no systematic change in the magnitude of GMST occurs when using this extra layer.

Using a three-layer EBM with longer calibration time series increases the estimate of peak warming under overshoot scenarios in all 15 ESMs tested compared to the use of a shorter experiment. This has policy relevance as the impacts of overshoot scenarios, and their compliance with global climate targets, depend on their level of peak warming (Riahi *et al* 2022). We also find that the year in which global warming thresholds are crossed is robustly brought a year or two sooner for policy-relevant GWLs.

Further work is needed to understand the mechanisms for, and full implications of, this dependence on the calibration length. While the stability of emulation parameters is improved when using 900 years, which we focus on due to the shrinking number of ESMs with available data beyond this length, individual ESMs show continued variation in their calibration for longer training periods. The results shown here use the individual EBM parameters taken from each ESM, projecting forward the associated responses using the EBM. One line of inquiry to explore is the connection to the 'hot model' problem; since the ESM-dependent projections warm when more years are available to calibrate the model, this discrepancy may be partially resolved using longer training data. It should also be noted that as models focused on global, smoothed temperatures, these EBMs cannot explicitly represent changes in higher-resolution processes such as ENSO under climate forcing.

Emulators play a key role in connecting diverse research areas, including between IPCC Working Groups (Kikstra *et al* 2022), and will continue to do so in the near future. It is imperative to better understand their behaviour, and how this connects to their calibration data. Using more EBM calibration data, and thereby better capturing the long-term response to forcing, produces climate responses to plausible scenarios on relevant timescales which are significantly warmer, with higher SLR on multi-century scales, all else being equal.

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.17950611.

Supplementary data available at https://doi.org/10.1088/1748-9326/ae3847/data1.

Code Availability

All code for analysis and production of all figures and tables can be found here: https://doi.org/10.5281/zenodo.17950611.

Acknowledgments

The authors acknowledge funding from the Horizon Europe projects WorldTrans and SPARCCLE (Grant Nos. 101081661 and 101081369).

ORCID iDs

Christopher D Wells 0000-0003-1958-0984
Donald P Cummins 0000-0003-3600-5367
Haozhe He 0000-0002-8168-4373
Chris Smith 0000-0003-0599-4633

References

Andrews T, Gregory J M, Paynter D, Silvers L G, Zhou C, Mauritsen T, Webb M J, Armour K C, Forster P M and Titchner H 2018 Accounting for changing temperature patterns increases historical estimates of climate sensitivity *Geophys. Res. Lett.* **45** 8490–9

Andrews T, Gregory J M and Webb M J 2015 The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models *J. Clim.* **28** 1630–48

Charney J G, Arakawa A, Baker D J, Bolin B, Dickinson R E, Goody R M, Leith C E, Stommel H M and Wunsch C I 1979 *Carbon Dioxide and Climate: A Scientific Assessment* (National Academy of Sciences)

Cummins D P, Stephenson D B and Stott P A 2020 Optimal estimation of stochastic energy balance model parameters *J. Clim.* **33** 7909–26

Dai A, Huang D, Rose B E J, Zhu J and Tian X 2020 Improved methods for estimating equilibrium climate sensitivity from transient warming simulations *Clim. Dyn.* **54** 4515–43

Dorheim K, Gering S, Gieseke R, Hartin C, Pressburger L, Shiklomanov A N, Smith S J, Tebaldi C, Woodard D L and Bond-Lamberty B 2024 Hector V3.2.0: functionality and performance of a reduced-complexity climate model *Geosci. Model Dev.* **17** 4855–69

Dunne J P *et al* 2025 An evolving Coupled Model Intercomparison Project Phase 7 (CMIP7) and Fast Track in support of future climate assessment *Geosci. Model Dev.* **18** 6671–700

Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization *Geosci. Model Dev.* **9** 1937–58

Flynn C M and Mauritsen T 2020 On the climate sensitivity and historical warming evolution in recent coupled model ensembles *Atmos. Chem. Phys.* **20** 7829–42

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 (Cambridge University Press) pp 923–1054

Forster P M *et al* 2025 Indicators of global climate change 2024: annual update of key indicators of the state of the climate system and human influence *Earth Syst. Sci. Data* **17** 2641–80

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 (Cambridge University Press) pp 1211–361

Fredriksen H-B, Eiselt K-U and Good P 2025 Exploring global temperature oscillations using a generalized linear box model *Geophys. Res. Lett.* **52** e2024GL113994

Fredriksen H-B and Rypdal M 2017 Long-range persistence in global surface temperatures explained by linear multibox energy balance models *J. Clim.* **30** 7157–68

Geoffroy O, Saint-Martin D, Bellon G, Voldoire A, Olivié D J L and Tytéca S 2013a Transient climate response in a two-layer energy-balance model. Part II: representation of the efficacy of deep-ocean heat uptake and validation for CMIP5 AOGCMs *J. Clim.* **26** 1859–76

Geoffroy O, Saint-Martin D, Olivié D J L, Voldoire A, Bellon G and Tytéca S 2013b Transient climate response in a two-layer energy-balance model. part i: analytical solution and parameter calibration using CMIP5 AOGCM experiments *J. Clim.* **26** 1841–57

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 A new method for diagnosing radiative forcing and climate sensitivity *Geophys. Res. Lett.* **31** L03205

Hausfather Z, Drake H F, Abbott T and Schmidt G A 2020 Evaluating the performance of past climate model projections *Geophys. Res. Lett.* **47** e2019GL085378

Held I M, Winton M, Takahashi K, Delworth T, Zeng F and Vallis G K 2010 Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing *J. Clim.* **23** 2418–27

Jackson L S, Maycock A C, Andrews T, Fredriksen H-B, Smith C J and Forster P M 2022 Errors in simple climate model emulations of past and future global temperature change *Geophys. Res. Lett.* **49** e2022GL098808

Jiménez-de-la-Cuesta D and Mauritsen T 2019 Emergent constraints on Earth's transient and equilibrium response to doubled CO_2 from post-1970s global warming *Nat. Geosci.* **12** 902–5

Kikstra J S *et al* 2022 The IPCC sixth assessment report WGIII climate assessment of mitigation pathways: from emissions to global temperatures *Geosci. Model Dev.* **15** 9075–109

Kopp R E *et al* 2023 The framework for assessing changes to sea-level (FACTS) v1.0: a platform for characterizing parametric and structural uncertainty in future global, relative, and extreme sea-level change *Geosci. Model Dev.* **16** 7461–89

Leach N J, Jenkins S, Nicholls Z, Smith C J, Lynch J, Cain M, Walsh T, Wu B, Tsutsui J and Allen M R 2021 FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration *Geosci. Model Dev.* **14** 3007–36

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 (Cambridge University Press) pp 553–672

Li S and Jarvis A 2009 Long run surface temperature dynamics of an A-OGCM: the HadCM3 $4\times CO_2$ forcing experiment revisited *Clim. Dyn.* **33** 817–25

Malagón-Santos V, Smith C, Fredriksen H-B, Hermans T H J, Edwards T L and Slangen A B A 2025 Emulating long-term CMIP6 projections of sterodynamic sea-level change using a three-layer energy balance model *Environ. Res. Lett.* **20** 84034

Meinshausen M *et al* 2020 The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500 *Geosci. Model Dev.* **13** 3571–605

Meinshausen M, Raper S C B and Wigley T M L 2011 Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6—part 1: model description and calibration *Atmos. Chem. Phys.* **11** 1417–56

Nicholls Z and Lewis J 2021 OpenSCM two layer model: a Python implementation of the two-layer climate model *J. Open Source Softw.* **6** 2766

Nordhaus W D 1991 To slow or not to slow: the economics of the greenhouse effect *Econ. J.* **101** 920–37

Ramme L, Blanz B, Wells C, Wong T E, Schoenberg W, Smith C and Li C 2025 Feedback-based sea level rise impact modelling for integrated assessment models with FRISIAv1.0 *Geosci. Model. Dev.* **18** 10017–52

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) (https://doi.org/10.1017/9781009157926.005)

Rugenstein M *et al* 2019 LongRunMIP: motivation and design for a large collection of millennial-length AOGCM simulations *Bull. Am. Meteorol. Soc.* **100** 2551–70

Rugenstein M *et al* 2020 Equilibrium climate sensitivity estimated by equilibrating climate models *Geophys. Res. Lett.* **47** e2019GL083898

Sandstad M, Aamaas B, Johansen A N, Lund M T, Peters G P, Samset B H, Sanderson B M and Skeie R B 2024 CICERO simple climate model (CICERO-SCM v1.1.1)—an improved simple climate model with a parameter calibration tool *Geosci. Model Dev.* **17** 6589–625

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 fair-calibrate v1.4.1: calibration, constraining, and validation of the FaIR simple climate model for reliable future climate projections *Geosci. Model Dev.* **17** 8569–92

Smith C *et al* 2021a *IPCC Working Group 1 (WG1) Sixth Assessment Report (AR6) Annex III Extended Data* (Zenodo) (https://doi.org/10.5281/zenodo.5705391)

Smith C, Nicholls Z R J, Armour K, Collins W, Forster P, Meinshausen M, Palmer M D and Watanabe M 2021b 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 1–35 (available at: www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter07_SM.pdf)

Smith C J and Forster P M 2021 Suppressed late-20th century warming in CMIP6 models explained by forcing and feedbacks *Geophys. Res. Lett.* **48** e2021GL094948

Smith C J *et al* 2020 Effective radiative forcing and adjustments in CMIP6 models *Atmos. Chem. Phys.* **20** 9591–618

Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98

Tebaldi C *et al* 2021 Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6 *Earth Syst. Dyn.* **12** 253–93

Tsutsui J 2017 Quantification of temperature response to CO_2 forcing in atmosphere–ocean general circulation models *Clim. Change* **140** 287–305

Tsutsui J and Smith C 2024 Revisiting two-layer energy balance models for climate assessment *Environ. Res. Lett.* **20** 14059

von Schuckmann K *et al* 2020 Heat stored in the Earth system: where does the energy go? *Earth Syst. Sci. Data* **12** 2013–41

Winton M, Takahashi K and Held I M 2010 Importance of ocean heat uptake efficacy to transient climate change *J. Clim.* **23** 2333–44

Yang H and Zhu J 2011 Equilibrium thermal response timescale of global oceans *Geophys. Res. Lett.* **38** L14711

Zehrung A, King A D, Nicholls Z, Zelinka M D and Meinshausen M 2025 Standardising the "Gregory method" for calculating equilibrium climate sensitivity *Geosci. Model Dev.* **18** 9433–50

Zelinka M D, Myers T A, McCoy D T, Po-Chedley S, Caldwell P M, Ceppi P, Klein S A and Taylor K E 2020 Causes of higher climate sensitivity in CMIP6 models *Geophys. Res. Lett.* **47** e2019GL085782

Wells C D, Cummins D P and Smith C 2025 ebm-calibration: revisions *Zenodo* (https://doi.org/10.5281/zenodo.17950611)