Diagnosing Stalled Warming in CMIP6 Models

Alyssa N. Poletti

A thesis
submitted in partial fulfillment of the
requirements
for the degree of

Master of Science

University of Washington
2022

Committee:
Dargan M. W. Frierson
Kyle C. Armour
Dennis L. Hartmann

Program Authorized to Offer Degree:
Department of Atmospheric Science

University of Washington

Abstract

Diagnosing Stalled Warming in CMIP6 Models

Alyssa N. Poletti

Chair of the Supervisory Committee:
Dargan M. W. Frierson
Department of Atmospheric Science

Global coupled models provide essential constraints on future Earth warming; however, significant spread across climate simulations creates uncertainty in the rate of warming. Nonlinearities in the rate of warming at different levels of CO2 contribute to spread even within a single model. We consider eight CMIP6 models for which both abrupt quadrupling (abrupt-4xCO2) and abrupt doubling (abrupt-2xCO2) of pre-Industrial CO2 experiments are available. Within these eight GCMs, two models have the largest difference in warming rate between abrupt-4xCO2 and abrupt-2xCO2: CESM2 and MRI-ESM2. Using estimated equilibrium statistics, a linearized energy balance model, and an analysis of ocean dynamics, we evaluate whether the net radiative feedback, the effective radiative forcing, or the effective heat capacity contribute to differences in the rate of warming across forcings and by what mechanism. We find that on average, radiative feedbacks are the dominant cause of forcing-dependent differences. However, for CESM2 and MRI-ESM2-0, the effective heat capacity is a significant, or even dominant, mechanism of slowed warming under abrupt doubling CO_2. We conclude that changes in the low cloud cover and depth of heat storage due to declines in the Atlantic Meridional Overturning Circulation (AMOC) play an important role in modulating the transient climate.

1. Introduction

The IPCC Sixth Assessment Report (AR6) projects our planet will warm between 2.1 to 3.5 degrees Celsius by 2100 under a "middle-of-the-road" emissions scenario: SSP2-4.5. Understanding how and at what rate the Earth System warms provides essential data for climate adaptation and mitigation. In this thesis, we will compare how and why the rate of transient warming varies across forcings, using an ensemble of general circulation models (GCMs) from the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al. 2016). Our ensemble is made up of eight GCMs for which experiments of both abrupt doubling (abrupt-2xCO2) and abrupt quadrupling (abrupt-4xCO2) from pre-industrial CO_2 levels are available.

In a comparison of the time series of globally average annual surface air temperature anomaly, each model shows more transient warming under CO_2 quadrupling than under CO_2 doubling, because the radiative forcing for 4xCO2 is approximately twice that of 2xCO2. But, surprisingly, the warming in abrupt-2xCO2 experiments is smaller compared to 1/2 times the warming in abrupt-4xCO2 experiments (Fig. 1a, b). Understanding the cause of this nonlinearity in transient warming, with relatively larger warming at higher CO_2 forcing, is integral to predictions of future warming since AR6 uses the latest generation of climate models, including these CMIP6 GCMs to constrain future warming scenarios. To better understand this nonlinearity, we will look at the two models with the largest difference in their transient warming across CO_2 forcing levels: CESM2 and MRI-ESM2-0.

To quantify warming differences across forcings, we consider a percent difference defined as:

$$= \frac{{}^{1}\!/_{2}\, T_{4x} - T_{2x}}{{}^{1}\!/_{2}\, T_{4x}}\ x\ 100\% \quad (1)$$

where T_{4x} and T_{2x} are the final temperatures in the abrupt-4xCO2 and abrupt-2xCO2 runs respectively, normalized for the forcing by dividing by two. When we compare the final decadal average of the temperature anomaly, we find that CESM2 and MRI-ESM2-0 are 32% and 35% colder in abrupt-2xCO2 than normalized (1/2 times) abrupt-4xCO2, while the average difference is 8% across the other six models. These two models are also outliers when we consider the rate of warming. Using a linear regression of the temperature anomaly over the years 30 to 150, most ensemble members have a smaller rate of warming in abrupt-2xCO2 compared to that in abrupt-4xCO2 (Table 1b). However, CESM2 and MRI-ESM-0 have the largest difference. More interestingly, MRI-ESM2-0 has a negative slope over those years. In other words, MRI-ESM2-0 has stalled in its warming over the final century of the 2xCO2 run (but, importantly, not in its 4xCO2 run).

To understand why CESM2 and MRI-ESM2-0 have significantly different transient behavior at different forcing levels, we will consider estimated equilibrium and time-dependent statistics. To quantify the transient warming, we will analyze the change in temperature from the years 30 to

150, the global mean temperature time series, and a linearized energy balance model (Donohoe et al. 2014):

$$C\,\frac{dT}{dt} = F + \lambda T(t) \quad (2)$$

where C is the heat capacity of the Earth system and t is time. In our energy balance equation, the time evolution of temperature is controlled by two variables: the heat capacity C and the radiative feedback parameter λ. The radiative feedback parameter (in units of W/m^2/K) encapsulates how components such as clouds, water vapor, albedo, and the lapse rate impact warming. The heat capacity (in units of J/m^2/K) represents the cumulative heat captured by our earth system, which is largely modulated by ocean dynamics.

The effective equilibrium climate sensitivity (ECS) is the final temperature of the Earth system once the planet reaches equilibrium after doubling pre-industrial levels of CO_2 (abrupt-2xCO2). ECS acts as an important measure of the planet's susceptibility to radiative forcings which constrains projections of future warming (Geoffroy & Saint-Martin 2020). The IPCC AR6 estimates that the planet's ECS is *likely* between 2.5 to 4 degrees Celsius with current emissions (IPCC 2021). Since the Earth system requires thousands of years to reach equilibrium, ECS is often estimated from much shorter GCM simulations using a "Gregory Regression": taking a linear regression of globally averaged surface air temperature anomaly and net top of atmosphere (TOA) radiative flux anomaly from a simulation where a constant radiative forcing is applied, such as abrupt doubling (abrupt-2xCO2) or quadrupling (abrupt-4xCO2) of CO_2 (Gregory et al. 2004):

$$R_{TOA}(t) = F + \lambda\, T(t) \quad (3)$$

where R is the TOA radiative anomaly, T is the surface air temperature anomaly, F is the forcing (or the effective radiative forcing; ERF), and λ is the radiative feedback parameter. The radiative feedback parameter represents the linearized response of top of atmosphere radiation per degree of warming. This approximation requires radiative feedback parameter to remain constant in time.

If the forcing is a doubling of CO_2, the ECS can be estimated as the x-intercept, or the temperature when the excess TOA radiative flux reaches zero:

$$ECS = T(R = 0) = -\frac{F}{\lambda} \quad (4)$$

ECS can be estimated using simulations with forcings other than a doubling of CO_2, provided the forcing is time invariant. In fact, simulations with an instantaneous quadrupling of CO_2 (abrupt-4xCO2) run for 150 years a common method to estimate the ECS, despite the fact that ECS is defined as the temperature response to doubling of CO_2. In an experiment with constant forcing F,

$$ECS = -\frac{F}{\lambda}\frac{F_{2x}}{F} = -\frac{F_{2x}}{\lambda} \quad (5)$$

where λ is estimated again by regression in outgoing longwave radiative-temperature (OLR-T) space. Assuming CO_2 forcing is logarithmic with concentration, $F = 2\,F_{2x}$ for an abrupt-4xCO2 simulation. If the OLR-T relation was truly linear, then ECS could be estimated with constant forcings of any magnitude, both positive and negative. However, analyses of CMIP5 and CMIP6 have found that the ECS estimated from abrupt-4xCO2 simulations are larger than those estimated from abrupt-2xCO2 simulations (Meraner et al. 2013, Good et al. 2015, Gregory et al. 2015, Mitevski et al. 2021).

We confirm that ECS estimated from abrupt-2xCO2 simulations is indeed smaller than ECS estimated from abrupt-4xCO2 simulations (Table 1a). Additionally, a global mean annual time series of surface air temperature anomaly reveals that on average the abrupt-4xCO2 transient warming is always larger than the abrupt-2xCO2 warming (Fig 1a, b). More interestingly, CESM2 and MRI-ESM2-0 also have the largest difference in effective ECS across forcings. A key question is whether differences in ECS alone are enough to explain the different character of warming under 2xCO2 compared to 4xCO2.

There are three likely components which may cause slowed or stalled warming in CESM2 and MRI-ESM2-0: the radiative feedback parameter, the effective radiative forcing, and the effective heat capacity. These components can lead to less warming in abrupt-2xCO2 experiments than in ½ abrupt-4xCO2 experiments through the following mechanisms: if the net radiative feedback is more negative in abrupt-2xCO2 experiments than in abrupt-4xCO2, the abrupt-2xCO2 effective radiative forcing is less than half that in abrupt-4xCO2, or if the effective heat capacity is larger in abrupt-2xCO2 than in abrupt-4xCO2. In this study, we will explore those three components through the following two research questions:

1. What are the roles of radiative feedbacks, effective radiative forcing, and effective heat capacity in driving slower/stalled warming in 2xCO2 relative to (1/2 times) 4xCO2?
2. What mechanisms cause differences in radiative feedbacks, effective radiative forcing, and effective heat capacity between 2xCO2 and 4xCO2 scenarios?

2. Data Availability, Resampling, Interpolation

This analysis is performed on CMIP6 model output data for which abrupt-4xCO2 and abrupt-2xCO2 data are available. The eight models are listed in table 1. We use monthly mean data resampled to annual means. The ensemble mean refers to an intermodel mean excluding CESM2 and MRI-ESM2-0, such that the two models of interest can be compared to the rest of the ensemble. When performing an ensemble mean across zonal or global structure, each model is interpolated to the CanESM5 grid. When calculating the change in feedbacks, we consider two timescales: the years 0 to 20 and the years 21 to 150, which we will call the fast and slow timescales (Armour 2017).

We calculate anomalies by subtracting the mean of the final 100 years of the pre-Industrial control (piControl) from the abrupt-4xCO2 or abrupt-2xCO2 data. We did not find significant drift in the last 100 years of the piControl surface air temperature. Therefore, we calculated the anomalies for atmospheric fields using a constant piControl value. The only model where the branch point from the parent simulation differs across forcings is CanESM5 (Table S1). Therefore, drift in the model energy balance would not significantly impact the results of this paper.

Atmospheric fields are three-dimensional (time, longitude, latitude) while ocean fields are four-dimensional (time, longitude, latitude, depth). For some models, the native grid for ocean fields is a tri-polar grid which we resample to a 1-degree by 1-degree lat-lon grid using bilinear interpolation with Climate Data Operators (CDO; Kasper et al. 2010). The command line operation is as follows:

cdo remapbil,r360x180 infile outfile

Atmospheric fields are available for our entire ensemble. The zonally averaged, ocean meridional overturning streamfunction is available for every model except GISS-E2-1-H and IPSL-CM6A-LR. Ocean potential temperature data is available for every model except MIROC6. See the supplemental table 1 and 2 for more details on data availability per model, model variants used, and model branch points.

3. Results & Methods

3.1. The roles of radiative feedbacks, effective radiative forcing, or effective heat capacity in causing slower/stalled warming in lower forcing scenarios

i. The Net Radiative Feedback across Forcings

Across models, differences in ECS increased from 2.1 to 4.7 K in CMIP5 to 1.8 to 5.6 K in CMIP6. Differences in ECS across CMIP6 models is largely explained by differences in the radiative feedback parameter (Dong et al. 2020, Zelinka et al. 2020). First, we will explore whether the radiative feedback parameter is different across forcing levels. Next, we will investigate *(i)* whether the radiative feedback changes differ across forcings, *(ii)* whether the radiative feedback changes over time, and whether any time dependence is also forcing-dependent.

A more negative radiative feedback parameter (lower ECS) in abrupt-2xCO2 than in abrupt-4xCO2 would contribute to slower/stalled warming in abrupt-2xCO2 than in abrupt-4xCO2. In table 2, we see that the radiative feedbacks across our ensemble are more negative in abrupt-2xCO2 compared to abrupt-4xCO2, except for GISS-E2-1-G. However, CESM2 and MRI-ESM2-0 are once again outliers, in that the radiative feedback parameter has the largest difference across forcings (Table 2a; Fig. 1c, d). CESM2 and MRI-ESM2-0 have a net radiative feedback parameter which is 0.33 and 0.58 $W/m^2/K$ less in abrupt-2xCO2 than abrupt-4xCO2, whereas the average difference is only 0.08 ± 0.07 $W/m^2/K$. The more negative abrupt-2xCO2 feedback parameter is consistent with the Gregory plots in figure 1c and d, where CESM2 and MRI-ESM2-0 both have a steeper slope and smaller ECS.

ii. The Net Radiative Feedback across Time

CMIP5 and CMIP6 models robustly show that the radiative feedback parameter becomes more positive (shallowing) over time following abrupt CO_2 forcing (Andrews et al. 2015; Dong et al. 2020). If the time dependence of the radiative feedback parameter is substantially smaller, or in the other direction (steepening), under abrupt-2xCO2 than abrupt-4xCO, then this difference in feedback behavior could contribute to slower/stalled warming under 2xCO2.

To understand the time dependence of the radiative feedback parameter, we consider the radiative feedbacks on a fast timescale (years 0 to 20), a slow timescale (years 21 to 150), and the difference between them. On average, the ensemble radiative feedback becomes more positive over time under both forcings, and the positive change is larger in abrupt-2xCO2 than abrupt-4xCO2 feedback (Table 2b; Fig. 1e, f). CESM2 is consistent with the ensemble; the radiative feedback becomes more positive under either forcing and when compared to the ensemble average. MRI-ESM2-0 also has a more positive radiative feedback over time, under both forcings, but that change is smaller in abrupt-2xCO2 than in abrupt-4xCO2. However, the time dependence of the MRI-ESM2-0 abrupt-2xCO2 radiative feedback is still within 1 standard

deviation of the ensemble average and is not the only model in the ensemble with a smaller change in abrupt-2xCO2 than abrupt-4xCO2.

If steepening radiative feedbacks are responsible for the stalled warming, then the difference between the climate feedback at the beginning of the run (years 0 to 20) and the end of the run (years 21 to 150) would be negative. This is not true for CESM2 and MRI-ESM2-0. In fact, only MIROC6 abrupt-2xCO2 sees a steepening radiative feedback. Thus, time-dependence of the net radiative feedback does not explain the stalled/slowed warming in CESM2 or MRI-ESM2-0.

iii. The Effective Radiative Forcing

A smaller effective radiative forcing (ERF) would lead to a smaller effective ECS and transient warming. So, if the CO_2 ERF under 2xCO2 (ERF_{2x}) was more than a factor of two smaller than the CO_2 ERF under 4xCO2 (ERF_{4x}), this could contribute to slower 2xCO2 warming compared to (1/2 times) 4xCO2. Additionally, to use the Gregory regression to calculate ECS_{4x}, we assume a logarithmic relationship between CO_2 and effective radiative forcing (ERF), where the ratio of ERF_{4x} and ERF_{2x} is two. Estimates of ERF which better match line-by-line radiative calculations find that the ratio of ERF_{4x} to ERF_{2x} is close to 2.09 (Etminan et al. 2016; Bryne & Goldblatt 2014). However, this ratio has variability across models. When fitting an energy balance model to CMIP6 models, Leach et al. (2020) found this ratio was 2.2 on average with a range of 1.9 to 2.5. Thus, differences in ERF across forcing levels may introduce differences in ECS and transient warming.

RFMIP was designed to calculate different models' radiative forcing with higher accuracy, but simulations only exist for 4xCO2 forcing, not 2xCO2 (Smith et al. 2020). Thus, we estimate ERF using a Gregory regression across the first 20 years of the 150-year abrupt 2xCO2 and 4xCO2 runs and interpolated to year 0 (Table 3). Within our ensemble, the average ratio of ERF_{4x} to ERF_{2x} is 2.039 +/- 0.18, meaning on average the effective radiative forcing is larger abrupt-4xCO2 than in abrupt-2xCO2. In CESM2 and MRI-ESM2-0, the ratios are 1.890 and 2.205 respectively. Therefore, effective radiative forcing cannot be a mechanism for suppressed warming in CESM2 as the ERF is proportionally larger in abrupt-2xCO2 than in abrupt-4xCO2. In MRI-ESM2-0, the ERF_{4x} is indeed proportionally larger than ERF_{2x}. However, the ERF_{4x}/ERF_{2x} ratio for MRI-ESM2-0 is within 1 standard deviation of the ensemble mean, which experiences only 8% difference between ½ abrupt-4xCO2 and abrupt-2xCO2 runs. Additionally, a ratio of 2.205 only corresponds to a ERF_{4x} 10% larger than twice ERF_{2x}, while MRI-ESM2-0 has a 35% difference in warming between ½ abrupt-4xCO2 and abrupt-2xCO2. Thus, proportionally smaller ERF_{2x} is not a plausible mechanism for suppressed warming in CESM2 and has only a small impact on slower/stalled warming in MRI-ESM2-0.

iv. The Effective Heat Capacity across Forcing

Earth system's thermal inertia makes ECS and the climate feedback parameter imperfect measures for understanding transient climate change. Our linearized energy balance model,

which provides a relationship between effective heat capacity and warming, has time dependence making it more representative of transient climate than ECS. Observations indicate that the ocean is the largest thermal reservoir, accounting for 84% of the increase in Earth's heat content from 1955 to 1998 (Levitus et al. 2005) and more than 90% of excess heat from anthropogenic warming (AR6). Therefore, our estimate of the Earth system's heat capacity is modulated by ocean dynamics. We can further conceptualize the effective heat capacity as an effective depth of heat storage in the ocean. Geoffroy et al. (2013) estimates the mean heat capacity of the deep ocean was 3.33 x 10^9 +/- 2.00 x 10^9 J/m^2/K for CMIP5 using the end of abrupt-4xCO2 simulations. While CMIP5 models have substantial differences in effective heat capacity across models, there isn't evidence that those difference impact transient warming (Geoffroy et al. 2012, Geoffroy et al. 2013), We will explore whether difference in the effective heat capacity across forcings can impact transient warming.

Equation 2 shows the typical form of our linearized energy balance model, which assumes a constant effective heat capacity. Without that assumption, the linearized energy balance model becomes:

$$\frac{d}{dt}\big(C(t) * T(t)\big) = F + \lambda T(t) = R_{toa} \quad (6)$$

In equation 6, we see that heat absorbed by the Earth system corresponds with either an increase in surface air temperature or in effective heat capacity. This formulation allows us to conceptualize how effective heat capacity changes with time in a GCM. We can calculate the heat capacity using the following equations (Donohoe et al. 2014):

$$C(t) = \frac{\int_0^t R_{toa}(t')\, dt'}{T(t)} \quad (7)$$

The time series of heat capacity for CESM2 and MRI-ESM2-0 differ from the ensemble mean (Fig 2a, b). For all models at any forcing level, the heat capacity monotonically increases in time. The ensemble mean for abrupt-2xCO2 and abrupt-4xCO2 are very similar, where the abrupt-2xCO2 effective heat capacity is slightly larger with more noise. In contrast, CESM2 and MRI-ESM2-0 have a significantly larger heat capacity in abrupt-2xCO2 versus abrupt-4xCO2.

In the abrupt-4xCO2 simulation, CESM2 and MRI-ESM2-0 are both within one standard deviation of the ensemble mean. In abrupt-2xCO2, CESM2 and MRI-ESM2-0 have an effective heat capacity which is not only larger but outside of one standard deviation of the ensemble mean. Recall, the ensemble mean excludes CESM2 and MRI-ESM2-0. At year 150, CESM2 abrupt-2xCO2 has an effective heat capacity which is approximately 0.86 x 10^9 J/m^2/K or almost 200 meters larger than abrupt-4xCO2. MRI-ESM2-0 has a difference of approximately 1.80 x 10^9 J/m^2/K or 400 meters. Therefore, CESM2 and MRI-ESM2-0 show significant differences in heat capacity across different forcings which is not present in our other models. This indicates

that we should look toward forcing-dependent differences in ocean dynamics to understand the mechanism responsible for slower/stalled warming in CESM2 and MRI-ESM2-0 under abrupt-2xCO2.

v. The Relative Importance of the Net Radiative Feedback and the Effective Heat Capacity

We found two plausible mechanisms for slower/stalled warming in CESM2 and MRI-ESM2-0 under abrupt-2xCO2: more-negative radiative feedbacks and/or a larger effective heat capacity. To explore the relative importance of each component, let's consider again the energy balance model. We can solve for the change in temperature:

$$\frac{dT}{dt} = \frac{F + \lambda T(t) - T\frac{dC}{dt}}{C} \quad (8)$$

Discretizing this formula, we get that:

$$\Delta T = \frac{\Delta t(F + \lambda T(t)) - T\Delta C}{C} \quad (9)$$

We can use our values for the radiative feedback parameter, the effective radiative forcing, and the effective heat capacity to recreate a global mean time series for each model at each forcing. To fit the temperature time series, we use a 1-year time step, and heat capacity data filtered using a one-degree polynomial fit with a window of 5 years to reduce the impact of natural variability in the TOA radiative flux anomaly (see supplemental table x for full EBM parameters). The EBM fits the global mean annual temperature data well with an average error of ±0.08 and ±0.14 K for the abrupt-4xCO2 and abrupt-2xCO2 ensemble mean data (Fig. 2e).

To understand the relative importance of radiative feedbacks and the effective heat capacity, we will run two experiments. First, we will run the EBM using the radiative feedback from abrupt-2xCO2 and the effective heat capacity from abrupt-4xCO2 (referred to as the "λ_{2x}/C_{4x}" experiment). Second, we will run the EBM using the radiative feedback from abrupt-4xCO2 and the effective heat capacity from abrupt-2xCO2 (referred to as the "λ_{4x}/C_{2x}" experiment). The time series of the temperature anomaly generated from the EBM along with experiments λ_{2x}/C_{4x} and λ_{4x}/C_{2x} are in figure 2c-e. Thus, the relative importance of the radiative feedback is,

$$= \frac{{}^{1}/_{2}\,T_{4x} - T_{\lambda 2x/C4x}}{{}^{1}/_{2}\,T_{4x} - T_{2x}}\, x\ 100\%, \quad (10)$$

and the relative importance of the effective heat capacity is,

$$= \frac{{}^{1}/_{2}\,T_{4x} - T_{\lambda 4x/C2x}}{{}^{1}/_{2}\,T_{4x} - T_{2x}}\, x\ 100\%, \quad (11)$$

with a residual of

$$= \left(1 - \frac{{}^1\!/_2\, T_{4x} - T_{\lambda 2x/C4x}}{{}^1\!/_2\, T_{4x} - T_{2x}} - \frac{{}^1\!/_2\, T_{4x} - T_{\lambda 4x/C2x}}{{}^1\!/_2\, T_{4x} - T_{2x}}\right) x\ 100\%, \quad (12)$$

Equations 10 through 12 are good approximations until,

$$\left({}^1\!/_2\, T_{4x} - T_{2x}\right) \to 0 \quad (13)$$

To estimate each relative importance, we will use the final decadal average of the global mean temperature time series. The relative importance of radiative feedbacks and the effective heat capacity for each model are shown in table 4. Note, GISS-E2-1-G is an edge case which is excluded from this analysis since the difference between the ½ abrupt-4xCO2 fit and the abrupt-2xCO2 fit is close to zero.

For CESM2, the radiative feedbacks account for 73% and the effective heat capacity account for 56% of the difference between ½ abrupt-4xCO2 and abrupt-2xCO2 runs (with a residual of -26%). In MRI-ESM2, radiative feedbacks account for 45% and the effective heat capacity accounts for 74% of that difference (with a residual of -19%). In comparison, radiative feedbacks account for the entire ½ abrupt-4xCO2 vs abrupt-2xCO2 warming difference (106%) in the rest of the ensemble. Additionally, the ensemble has residuals less than 5%, while CESM2 and MRI-ESM2 have large residuals.

Therefore, forcing-dependent differences in radiative feedbacks are the dominant mechanism of slowed warming in CESM2, while effective heat capacity is the dominant mechanism of slowed warming in MRI-ESM2-0. This is consistent with the previous section as MRI-ESM2-0 has the larger effective heat capacity, while CESM2 has a more negative radiative feedback parameter in abrupt-2xCO2 than MRI-ESM2-0. Additionally, effective heat capacity is a significant, but secondary, mechanism of slowed warming in CESM2 when compared to the rest of the ensemble. Effective heat capacity only accounts for -6% of the difference in warming in the other six models. While difference in the radiative feedback parameter is the dominant mechanism of differences in warming across models, we see here that the effective heat capacity can be a significant component in forcing-dependent differences.

3.2. What mechanism leads to differences in radiative feedbacks and effective heat capacity in lower forcing scenarios?

i. Feedback Decomposition

To understand why the radiative feedback parameter in CESM2 and MRI-ESM2-0 is more negative in abrupt-2xCO2 simulations, we will consider a decomposition of the climate feedbacks. The climate feedback parameter is the sum of each different radiative feedback, such as the Planck feedback (*B*), cloud feedbacks (*CRE*), the water vapor feedback (*WV*), the lapse rate feedback (*LR*), the ice-albedo feedback (*A*), etc.:

$$\lambda = \sum_J \lambda_j = \lambda_{CRE} + \lambda_{Clear\ Sky} = (\lambda_{SW\ CRE} + \lambda_{LW\ CRE}) + (\lambda_B + \lambda_{WV} + \lambda_A + \lambda_{LR} + \cdots) \quad (14)$$

We decompose the radiative feedbacks by applying a Gregory regression to clear sky and cloudy radiative fields, which are two-dimensional model outputs. Using this decomposition, we can estimate the change in the radiative feedbacks for net, shortwave, and longwave CRE as well as net, shortwave, and longwave clear sky feedbacks. This method is further described in Andrews et al. (2012).

When we perform this decomposition on CESM2 and MRI-ESM2-0, we find that shortwave cloud feedbacks are the dominant mechanism through which the abrupt-2xCO2 radiative feedback steepens. In table 5a, the shortwave cloud radiative effect, the longwave cloud radiative effect, and the longwave clear sky feedback are all more negative in abrupt-2xCO2 than abrupt-4xCO2 in CESM2. In MRI-ESM2-0, the shortwave cloud radiative effect is more negative in abrupt-2xCO2 than in abrupt-4xCO2. The longwave clear sky effect is also more negative in the lower forcing scenario but is less than shortwave cloud changes. It's consistent with the previous results that CESM2 has more components which are more negative in abrupt-2xCO2 since radiative feedback changes are the dominant mechanism through which CESM2 has slowed warming in abrupt-2xCO2.

ii. Shortwave Cloud Radiative Effect

Since the shortwave cloud radiative effect has the largest difference across forcings for both CESM2 and MRI-ESM2-0, we will further investigate the radiative effects using the approximate partial radiative (APRP) method described in Taylor et al. (2007) over our ensemble. APRP only consider shortwave effects which can be broken down into cloud, clear sky, and surface albedo responses. Thus, analysis of longwave CRE and clear sky effects in CESM2 will have to be investigated in future research using a radiative kernel method.

Figure 3a shows the difference (abrupt-2xCO2 – ½ abrupt-4xCO2) in the shortwave cloud radiative effect. Therefore, negative values represent locations where the shortwave CRE is smaller in abrupt-2xCO2 simulations. In the ensemble, excluding CESM2 and MRI-ESM2-0, the shortwave CRE is more positive in abrupt-2xCO2. In contrast, CESM2 and MRI-ESM2-0 has a strongly negative signal in the North Atlantic, south of Greenland. This region of Atlantic enhanced low cloud cover corresponds with the North Atlantic warming hole (NAWH; Fig. 3b). This region of depressed sea surface temperatures (SSTs) and enhanced low cloud cover is indicative of a decline in the Atlantic Meridional Overturning Circulation (AMOC; Zhang et al. 2010, Jackson et al. 2015).

The AMOC is a meridional circulation which transports heat from the South Atlantic to the North Atlantic. The AMOC strength tends to decrease under warming, though estimates of AMOC decline introduce significant variability across model-space (Lin et al. 2019). Lin et al. (2019) suggests that hemispheric temperature asymmetry, caused by changes in meridional ocean heat transport, is the main uncertainty in the evolution of the surface warming pattern across models. Changes to the spatial pattern of warming due to AMOC decline have impacts on radiative feedbacks. In more detail, AMOC decline leads to cold SSTs, increased estimated

inversion strength, and enhanced low cloud cover in the North Atlantic (Drijfhout et al. 2012, Rugenstein et al. 2013, Drijfhout 2015, Trossman et al. 2016, Lin et al. 2019, Hu et al. 2020), just as we see in CESM2 and MRI-ESM2-0 in abrupt-2xCO2.

iii. Atlantic Meridional Overturning Circulation

To understand how changes in ocean dynamics impact the surface warming pattern, radiative feedbacks, and the effective heat capacity, we will analyze the AMOC strength anomaly, the northward ocean heat transport, and the effective depth of heat storage. First, we will consider differences in the magnitude of AMOC decline across models.

Not only is AMOC decline a mechanism for negative, shortwave cloud radiative effects, there is substantial evidence that AMOC decline decreases the transient climate response particularly in the Northern Hemisphere (Winton et al. 2013, Rugenstein et al. 2013, Drijfhout 2015, Good et al. 2015, Hu et al. 2015, Trossman et al. 2016, Lin et al. 2019, Hu et al. 2020, Bonnet et al. 2021, Liu et al. 2021). In ECHAM5/MPI-OM GCM, abrupt AMOC shutdown led to cooling which competed with global warming (Drijfhout 2015). Additionally, ECHAM/MIP-OM exhibited hemispheric asymmetry in warming similar to CESM2 (Fig. 3).

Drijfhout (2015) attributed this slowed warming to a decrease in cross-equatorial heat transport by AMOC of 0.6×10^{15} W, which is equivalent to an increase in ocean heat uptake or decrease in surface forcing of 2 W/m^2. In HadGEM2-E2, the AMOC declines 35% more under doubled CO_2 than quadrupled CO_2 (Good et al. 2015). Since the impact that AMOC decline has on transient climate is proportional to the magnitude of the decline (Hu et al. 2020), this may indicate that enhanced AMOC decline in lower forcing scenarios introduces a fractionally larger decrease in transient climate in abrupt-2xCO2 than abrupt-4xCO2 due to a lower limit in AMOC strength. In other words, some models approach nearly complete AMOC shut down under abrupt-2xCO2, leaving little room for further decrease in abrupt-4xCO2.

We define the AMOC strength as the maximum streamfunction anomaly value north of 30N and below 500 meters (Lin et al. 2019) and the AMOC decline as the difference between the first five and last five years. We chose a 5-year window since the change in AMOC strength in the first 20 years is drastic (Fig. 4a). In abrupt-4xCO2, the ensemble average AMOC strength decreases while in abrupt-2xCO2, the ensemble average AMOC strength decreases and then recovers. The abrupt-2xCO2 ensemble average has larger differences across models, due to some models experiencing AMOC decline and some experiencing an AMOC decline then recovery.

CESM2 and MRI-ESM2-0 experience fractionally larger AMOC declines in abrupt-2xCO2 than in abrupt-4xCO2 (Fig. 4a). As in, the abrupt-2xCO2 AMOC decline is larger than ½ the abrupt-4xCO2 decline. CESM2 and MRI-ESM2-0 decline 72% and 94% of the abrupt-4xCO2 value under half of the forcing (Table 6). Thus, the two models with the largest AMOC decline experience the largest increase in North Atlantic low cloud cover, more negative radiative feedbacks, and the largest increases in effective heat capacity in the lower forcing scenario.

iv. Ocean Heat Content and Transport

Clearly, ocean dynamics play an important role in suppressing warming in both CESM2 and MRI-ESM2 in abrupt-2xCO2 simulations through sea surface temperatures and radiative feedbacks. However, the effective heat capacity is also impacted by ocean dynamics as it represents the effective depth of heat storage. Therefore, we will analyze how ocean heat transport (OHT) and content (OHC) change in CESM2 and MRI-ESM2-0.

To find the ocean heat transport, we have to first calculate the ocean heat content and surface heat flux. The surface flux (Q) is the sum of the following surface radiative terms:

$$Q(t,\lambda,\theta) = SW^{\downarrow} + LW^{\downarrow} - SW^{\uparrow} - LW^{\uparrow} - LH - SH - L_f \, x \, SN \quad (15)$$

Where t is time, λ is latitude, θ is longitude, *SW* is shortwave radiative flux, *LW* is longwave radiative flux, the up arrow denotes upward flux, the down arrow denotes downward flux, *LH* is latent heat flux, *SH* is sensible heat flux, L_f is the latent heat of fusion, and *SN* is the snow melt flux.

We can take the zonal mean of the net surface flux by:

$$\overline{Q}(t,\lambda) = \int_0^{2\pi} Q(t,\lambda,\theta) d\theta \quad (16)$$

We will denote zonal means with the bar. The ocean heat content (OHC) is defined as:

$$OHC(t,z,\lambda,\theta) = \rho C_p T(t,z,\lambda,\theta) \quad (17)$$

Where z is depth, ρ is the density of sea water (1030 kg/m^3), C_p is the specific heat of water under constant pressure (4184 J/Kg/K), *A* is the area of the grid cell, and *T* is the potential temperature of ocean grid cell with depth. The density and specific heat convert the temperature to an energy term. The quotient of the area and the surface area of the planet provides an area average at each grid cell. The column integrated heat content is:

$$OHC(t,\lambda,\theta) = \int_0^{Z_B} \rho C_p T(t,z,\lambda,\theta)\, dz \quad (18)$$

The zonal mean of the column integrated heat content is:

$$\overline{OHC}(t,\lambda) = \int_0^{2\pi} \int_0^{Z_B} \rho C_p T(t,z,\lambda,\theta)\, dz\, d\theta \quad (19)$$

The net surface heat flux is in W/m^2, while the ocean heat content is in J/m^2. To compare these terms, we have to consider the time derivative of ocean heat content:

$$\frac{\partial}{\partial t}\overline{OHC}(t,\lambda) = \frac{\partial}{\partial t}\int_0^{2\pi}\int_0^{Z_B} \rho C_p T(t,z,\lambda,\theta)\, dz\, d\theta \quad (20)$$

The ocean heat transport is defined as the meridional integral of the difference between surface heat flux and the time derivative of the ocean heat content. We will call the integrand the ocean heat divergence:

$$OHT = \int_{\lambda}^{90^o N} \frac{A(\lambda,\theta)}{\int_\lambda \int_0^{2\pi} A\, d\lambda\, d\theta}\, \overline{OHD}\, d\lambda \quad (21)$$

$$OHD = Q - \frac{\partial}{\partial t} OHC \quad (22)$$

Where the first term in equation 21 is an area weighting. The zonal mean of ocean heat divergence is:

$$\overline{OHD}(t,\lambda) = \int_0^{2\pi} \left(Q - \frac{\partial}{\partial t} OHC\right) d\theta \quad (23)$$

We can rearrange equation 23:

$$\overline{OHD}(t,\lambda) = \int_0^{2\pi} Q\, d\theta - \frac{\partial}{\partial t}\int_0^{2\pi} OHC\, d\theta \quad (24)$$

$$\overline{OHD}(t,\lambda) = \overline{Q}(t,\lambda) - \frac{\partial}{\partial t}\overline{OHC}(t,\lambda) \quad (25)$$

Therefore, the northward ocean heat transport becomes:

$$OHT = \int_{\lambda}^{90^o N} \frac{A(\lambda,\theta)}{\int_\lambda \int_0^{2\pi} A\, d\lambda\, d\theta}\left(\overline{Q}(t,\lambda) - \frac{\partial}{\partial t}\overline{OHC}(t,\lambda)\right) d\lambda \quad (26)$$

Combining equations 26, 16, and 20, we can represent the northward ocean heat transport in terms of net surface flux and potential temperature:

$$OHT = \int_{\lambda}^{90^o N} \frac{A(\lambda,\theta)}{\int_\lambda \int_0^{2\pi} A\, d\lambda\, d\theta}\left[\left(\int_\theta^{2\pi} Q d\theta\right) - \left(\frac{\partial}{\partial t}\int_\theta^{2\pi}\int_0^{Z_B} \rho C_p T(t,z,\lambda,\theta)\, dz d\theta\right)\right] d\lambda \quad (27)$$

Now, we can calculate the ocean heat content and the northward ocean heat transport using potential temperature. On average, the ensemble sees a decrease in northward heat transport in

the Northern midlatitudes (Fig. 4b). However, CESM2 and MRI-ESM2-0 both have the largest decrease in transport and a proportionally larger decline in abrupt-2xCO2, in the Northern midlatitudes. This decrease in Northern midlatitude heat transport provides additionally support that CESM2 and MRI-ESM2-0 are outliers in the magnitude of AMOC decline within the ensemble and also experience variability in AMOC decline across forcings.

Next, let's consider the ocean heat content. The effective heat capacity can be conceptualized as an effective depth of heat storage. Thus, we can use our analysis of column integrated heat storage to find the depth at which this heat is stored:

$$Percent\ of\ Heat\ Storage(z) = \frac{\int_0^Z \rho C_p T(t, z', \lambda, \theta)\, dz'}{\int_0^{Z_B} \rho C_p T(t, z', \lambda, \theta)\, dz'}\ x\ 100\% \quad (28)$$

Using this formula, we can solve for some depth *z* at which some percentage of heat storage has accumulated. Thus, we can choose some percent to represent our depth of heat storage. We performed a linear regression without a *y*-intercept of the final effective depth (calculated from our effective heat capacity) and the depth of percent heat storage, to determine which depth of percent heat storage best represents the effective depth. The depth of 55% heat storage has both a slope close to one and a *p*-value less than 0.05 (Fig 5c); however, these results are robust between 50% and 80% depth of heat storage.

Figure 5a shows the globally averaged depth of 55% heat storage over time. On average, the ensemble always has deeper heat penetration in abrupt-2xCO2 than in abrupt-4xCO2. This is also true for CESM2 and MRI-ESM2-0. For both simulations, MRI-ESM2-0 is on the upper end of the 55% heat content depth uncertainty, indicating that MRI-ESM2-0 has deeper heat penetration than the ensemble at either forcing. On the other hand, CESM2 is at the lower end of variability.

When we consider the difference across forcings (Fig. 5b), we see that the difference between abrupt-2xCO2 and abrupt-4xCO2 for the ensemble increases in time. Though CESM2 has a shallower than average depth of heat storage, the difference between the abrupt-2xCO2 and abrupt-4xCO2 depth of heat storage is larger in CESM2 than the ensemble mean. In MRI-ESM2-0, the difference in depth of heat storage across forcings increases for the first 20 years and then begins to plateau after year 40. The difference in depth of heat storage in MRI-ESM2-0 is larger than the ensemble until year 120. Additionally, the plateau in the difference in depth of heat storage coincides with the stalled warming.

Next, we can consider the zonal structure of the depth of 55% heat storage, to determine which basins are primarily responsible for the enhanced depth of heat storage in CESM2 and MRI-ESM2-0 abrupt-2xCO2. We use the same formulae to determine the zonal depth of heat storage as the globally averaged depth of heat storage, by omitting the latitudinal integral. For the zonal structure, we omit values of the temperature anomaly which are less than 0 to avoid values where

the percent of heat storage are less than 0. The results without omitting negative temperature anomalies are robust but have more noise.

Figure 6a shows the difference in the zonal depth of 55% heat storage between abrupt-2xCO2 and abrupt-4xCO2. All of the models show larger peaks in depth of heat storage in the Southern Ocean and the North Atlantic/North Pacific in abrupt-2xCO2 simulations. CESM2 has enhanced depth of heat storage in abrupt-2xCO2 at all latitudes, compared to the ensemble mean. The difference in depth of heat storage for MRI-ESM2-0 is larger than the ensemble mean only in the North Atlantic/North Pacific. The zonal structure of the ocean potential temperature is consistent with the depth of heat storage since CESM2 and MRI-ESM2-0 show cooler temperatures at depth in the North Atlantic/North Pacific than the ensemble mean. As the AMOC is a substantial mechanism of deep water formation in the Northern ocean basin, this may indicate that the fractionally larger AMOC decline in MRI-ESM2-0 abrupt-2xCO2 enhances the depth of heat storage through enhanced stratification (Winton et al. 2013, Kostov et al. 2014).

4. Discussion

We began this study with the observation that the transient climate and estimated equilibrium climate sensitivity in CESM2 and MRI-ESM2 was significantly different in abrupt-2xCO2 than in abrupt-4xCO2. While ECS in abrupt-4xCO2 is consistently larger than in abrupt-2xCO2, after adjusting for forcing, the abrupt-2xCO2 ECS in CESM2 and MRI-ESM2-0 had the largest difference from abrupt-4xCO2. Additionally, these models showed stalled and slowed warming in the transient climate of abrupt-2xCO2 simulations in contrast to monotonically increasing warming in the ensemble mean. Thus, we asked these two guiding questions:

1. What are the roles of radiative feedbacks, effective radiative forcing, and effective heat capacity in driving slower/stalled warming in 2xCO2 relative to (1/2 times) 4xCO2?
2. What mechanisms cause differences in radiative feedbacks, effective radiative forcing, and effective heat capacity between 2xCO2 and 4xCO2 scenarios?

In investigating the radiative feedbacks, the effective radiative forcing, and the effective heat capacity of our ensemble, we found that differences in the radiative feedbacks and the effective heat capacity across forcings are the two dominant mechanisms of the slowed/stalled warming in CESM2 and MRI-ESM2-0 under an abrupt doubling of CO_2. With a linearized energy balance model, we determined that radiative feedbacks are responsible for 70% and 45% of the difference in final warming in CESM2 and MRI-ESM2-0. The effective heat capacity was responsible for 56% and 74% of that difference in CESM2 and MRI-ESM2-0, with a residual of -26% and -19%. Essentially, these models largely differ from the ensemble in two ways: the larger feedback differences across forcings and the larger role of effective heat capacity in modulating the rate of warming.

i. Radiative Feedbacks

We explored how the radiative feedbacks change across forcings and across time. We found that more negative radiative feedbacks in abrupt-2xCO2 than in abrupt-4xCO2 led to a smaller ECS_{2x} in CESM2 and MRI-ESM2-0. In contrast, the evolution of the radiative feedback parameter with time would not lead to the smaller ECS seen in abrupt-2xCO2 scenarios for CESM2 and MRI-ESM2-0. Finally, the difference in effective radiative forcing across forcings is not significant enough to cause the dramatic difference in warming in our two models.

In a decomposition of the radiative feedbacks, we found that the shortwave cloud radiative effect had the largest, negative difference between abrupt-4xCO2 and abrupt-2xCO2 for both models. For CESM2, the shortwave clear sky and longwave cloud radiative effect also lead to a more negative radiative feedback in the lower forcing scenario. Using an approximate radiative perturbation method, we found that the decrease in the shortwave cloud feedback is due to enhanced cloud cover in the North Atlantic, indicative of decline in the Atlantic Meridional Overturning Circulation. An analysis of the AMOC strength confirms that CESM2 and MRI-ESM2-0 indeed experience a proportionally larger AMOC decline in abrupt-2xCO2 than in abrupt-4xCO2. Thus, nonlinearities in how Atlantic Ocean circulations change are the dominant

mechanism of more negative radiative feedbacks and a significant mechanism of the slowed/stalled warming in CESM2 and MRI-ESM2-0 under an abrupt doubling of CO2.

ii. Effective Heat Capacity

The effective heat capacity is the dominant mechanism of stalled warming in MRI-ESM2-0 and a significant mechanism of slowed warming in CESM2. In an analysis of differences in effective heat capacity across forcings, we found there is virtually no variability across forcing for the ensemble, except for CESM2 and MRI-ESM2-0 where the heat capacity increases dramatically in time under abrupt-2xCO2.

Next, we explore the mechanism of heat capacity changes. Since the effective heat capacity is analogous to an effective depth of heat storage, we consider the global mean and zonal mean heat storage against ocean depth. An analysis of the depth at which the ocean contains 55% of the total column integrated heat, we found that MRI-ESM2-0 and CESM2 increase in effective depth more rapidly than the ensemble.

An analysis of the zonal ocean heat storage with depth reveals that all models have a peak in depth of heat storage in the Southern Ocean and North Atlantic which is larger in abrupt-2xCO2 than abrupt-4xCO2. This may be due to increased stratification in abrupt-2xCO2. The stratification still exists in abrupt-4xCO2 but enhanced surface warming shallows the depth of heat storage more than stratification deepens it. For CESM2, the difference in depth of heat storage across forcings is larger than the ensemble mean for all latitudes, while in MRI-ESM2-0- this is only true for the North Atlantic. Since we discovered that heat capacity is the dominant mechanism of stalled warming in MRI-ESM2-0 and that MRI-ESM2-0 has a fractionally larger AMOC decline in abrupt-2xCO2, this may indicate that AMOC decline can increase the effective heat capacity by enhancing the depth of heat storage in the North Atlantic. This may also be true for CESM2, but to a lesser extent since the radiative feedback parameter is the dominant mechanism of slowed warming in abrupt-2xCO2.

5. Conclusion

While it is expected from previous literature that both the transient climate and the effective ECS is disproportionately larger in abrupt-4xCO2 simulations than in abrupt-2xCO2 simulations, previous literature attributes the difference in ECS across models and across forcings to radiative feedbacks (Meraner et al. 2013, Good et al. 2015, Gregory et al. 2015, Mitevski et al. 2021). In our eight-model ensemble, we confirmed that differences in warming between abrupt-4xCO2 and abrupt-2xCO2 simulations is largely due to the radiative feedback parameter. However, in this study we found that in models with the largest difference in warming between abrupt-4xCO2 and abrupt-2xCO2 the difference is caused by both radiative feedbacks *and* the effective heat capacity. We reached this conclusion through an analysis of the slowed/stalled warming CESM2 and MRI-ESM2 in comparison to the rest of our ensemble.

For CESM2 and MRI-ESM2, the more negative radiative feedback parameter in abrupt-2xCO2 is in part caused by depressed shortwave cloud feedbacks due to variability in AMOC decline across forcings. Changes in the longwave clear sky and cloud feedbacks may also play a part in the slowed warming in CESM2. APRP analysis revealed that the large contributor to the more negative shortwave cloud feedback is enhanced low cloud over in the North Atlantic, likely due to AMOC decline (Zhang et al. 2010, Jackson et al. 2015). While our analysis is limited to shortwave effects using the APRP method, future work can consider longwave feedback changes in more detail using radiative kernel methods.

In these two models, the larger heat capacity in abrupt-2xCO2 is due to enhanced depth of heat storage, particularly in the Southern Ocean and Atlantic Ocean. This may be due to enhanced stratification in abrupt-2xCO2 runs due in part to fractionally larger AMOC decline in the abrupt-2xCO2 simulations for CESM2 and MRI-ESM2-0. In abrupt-4xCO2 simulations, the impact of enhanced stratification is outweighed by substantial surface warming. This study does not investigate ocean potential density and salinity profiles. Thus, future work could identify the mechanism through which AMOC decline increases stratification, as well as why the Southern Ocean sees an increase in depth of heat storage in lower forcing simulations.

Ultimately, nonlinearities in the magnitude of AMOC decline across forcings is a likely cause of slowed/stalled warming in CESM2 and MRI-ESM2-0 under an abrupt doubling of CO_2, through both radiative feedbacks and the effective heat capacity. The phenomenon where AMOC decline is proportionally larger in abrupt-2xCO2 (Good et al. 2016) or abrupt-3xCO2 (Mitevski et al. 2021) is well documented but not well understood. If the AMOC declines by more than 50% of its initial strength in abrupt-2xCO2 simulations, then a linear response to abrupt-4xCO2 is impossible unless the circulation is to reverse. Future research could investigate how the base state, the physics, or internal variability of models impacts differences in the magnitude of AMOC decline across forcings. However, this analysis makes clear that forcing-dependent changes ocean dynamics can dramatically impact the rate of transient warming across GCMs.

Additionally, this study is constrained by 150-year runs while interactions with the intermediate and deep ocean occur on centennial to millennial timescales. Model runtimes beyond 150 years would aid in understanding how AMOC recovery impacts transient climate. The magnitude of AMOC decline is largely uncertain in GCMs and our understanding of AMOC evolution is constrained by our run time since AMOC recovery which may occur on centennial timescales (Weaver et al. 2012, Cheng et al. 2013, Ackermann et al. 2020). Without statistically significant observational data on trends in AMOC strength (Lobelle et al. 2020, Jackson et al. 2020), we rely on GCMs for predicting how AMOC will change. In a model with large AMOC decline and suppressed global warming, the subsequent AMOC recovery could lead to a period of enhanced warming. Thus, model runs beyond 150-years would allow for a continued analysis of the impact of Atlantic and Southern Ocean dynamics on the transient climate. Efforts like LongrunMIP are critical to improving understanding of longer-term transient climate across many models (Rugenstein et al. 2019).

This study is further constrained by the number of available abrupt-2xCO2 datasets. A larger sample of abrupt-2xCO2 simulations will allow future work to better understand how radiative feedbacks and the effective heat capacity impact transient warming. The connection between CO_2 forcing and ocean stratification may have a larger impact on transient warming than previously thought. As we exceed 1.5 times preindustrial CO_2 concentrations (UK Mett Office 2021), simulations of transient climate and ocean dynamics in abrupt-2xCO2 simulations become increasingly relevant for projecting global warming in the 21st century.

6. Tables & Figures

	(a) Equilibrium Response			
Model	**ECS_{4xCO2}** [K]	**ECS_{2xCO2}** [K]	**Difference** [K]	**% Difference**
CanESM5	5.67	4.67	1.00	18%
CESM2	**5.24**	**3.37**	**1.90**	**36%**
CNRM-CM6-1	4.86	4.20	0.66	14%
GISS-E2-1-G	2.70	2.63	0.07	2.6%
GISS-E2-1-H	3.10	2.96	0.14	4.5%
IPSL-CM6A-LR	4.59	3.97	0.62	13%
MIROC6	2.54	2.18	0.36	14%
MRI-ESM2-0	**3.13**	**2.48**	**0.65**	**21%**
Ensemble Mean	*3.91 ± 1.30*	*3.44 ± 0.98*	*0.48 ± 0.35*	*11% ± 6%*

	(b) Transient Response			
Model	**dT/dt_{4xCO2}** [10^{-3} K/yr]	**dT/dt_{2xCO2}** [10^{-3} K/yr]	**Difference** [10^{-3} K/yr]	**% Difference**
CanESM5	7.58	5.82	1.76	23%
CESM2	**9.74**	**3.90**	**5.84**	**60%**
CNRM-CM6-1	10.05	7.74	2.31	23%
GISS-E2-1-G	5.48	4.42	1.06	19%
GISS-E2-1-H	7.56	3.54	4.02	53%
IPSL-CM6A-LR	11.70	5.71	5.99	51%
MIROC6	1.72	3.00	-1.28	-74%
MRI-ESM2-0	**11.30**	**-0.50**	**11.8**	**104%**
Ensemble Mean	*7.35 ± 3.51*	*5.04 ± 1.74*	*2.31 ± 2.50*	*16% ± 48%*

Table 1: (a) The effective equilibrium climate sensitivity estimated from abrupt-4xCO2 and abrupt-2xCO2, the difference, and the percent difference across forcings. (b) The rate of change of the temperature for the years 30 to 150, the difference, and the percent difference across forcings, calculated using a calculated using a linear regression of annually averaged temperature anomaly versus time. The difference is calculated for both (a) and (b) as ½ abrupt-4xCO2 minus abrupt-2xCO2. The ensemble mean excludes CESM2 and MRI-ESM2-0.

	(a) Net Radiative Feedback Parameter		
	Abrupt-4xCO2	**Abrupt-2xCO2**	**Difference**
Model	**λ (Years 0 to 150)**	**λ (Years 0 to 150)**	**$\lambda_{2x} - \lambda_{4x}$**
CanESM5	-0.66	-0.76	-0.10
CESM2	**-0.63**	**-1.21**	**-0.58**
CNRM-CM6-1	-0.73	-0.78	-0.05
GISS-E2-1-G	-1.44	-1.43	0.01
GISS-E2-1-H	-1.17	-1.27	-0.10
IPSL-CM6A-LR	-0.77	-0.93	-0.16
MIROC6	-1.46	-1.62	-0.17
MRI-ESM2-0	**-1.07**	**-1.40**	**-0.33**
Ensemble Mean	*-1.04 ± 0.37*	*-1.13 ± 0.36*	*-0.08 ± 0.07*

Table 2: (a) The net radiative feedback parameter in abrupt-4xCO2, abrupt-2xCO2, and the difference across forcings. The ensemble mean excludes CESM2 and MRI-ESM2-0.

	(b) Radiative Feedback Parameter on fast and slow timescales					
	Abrupt-4xCO2 [w/m^2/K]			**Abrupt-2xCO2** [w/m^2/K]		
Model	**λf (Years 0 to 20)**	**λs (Years 21 to 150)**	**Difference (λs – λf)**	**λf (Years 0 to 20)**	**λs (Years 21 to 150)**	**Difference (λs – λf)**
CanESM5	-0.68	-0.61	0.07	-0.87	-0.64	0.23
CESM2	**-1.14**	**-0.41**	**0.73**	**-1.57**	**-0.99**	**0.58**
CNRM-CM6-1	-0.90	-0.77	0.13	-1.01	-0.59	0.42
GISS-E2-1-G	-1.42	-1.22	0.20	-1.66	-1.31	0.35
GISS-E2-1-H	-1.27	-1.16	0.11	-1.43	-1.29	0.14
IPSL-CM6A-LR	-1.00	-0.64	0.36	-1.10	-0.74	0.36
MIROC6	-1.63	-1.63	0.00	-1.40	-1.76	-0.36
MRI-ESM2-0	**-1.33**	**-0.80**	**0.53**	**-1.26**	**-1.15**	**0.11**
Ensemble Mean	*-1.15 ± 0.35*	*-1.01 ± 0.40*	*0.15 ± 0.12*	*-1.25 ± 0.30*	*-1.06 ± 0.48*	*0.19 ± 0.29*

(b) The net radiative feedback parameter over the fast timescale (years 0 to 20), the slow timescale (years 21 to 150), and the difference (slow minus fast) for both abrupt-4xCO2 and abrupt-2xCO2. The ensemble mean excludes CESM2 and MRI-ESM2-0

	Effective Radiative Forcing		
	Abrupt-2xCO2	**Abrupt-4xCO2**	**Ratio**
Model	**ERF_{2x}**	**ERF_{4x}**	**ERF_{4x}/ERF_{2x}**
CanESM5	3.775	7.602	2.014
CESM2	**4.546**	**8.596**	**1.890**
CNRM-CM6-1	3.761	7.670	2.040
GISS-E2-1-G	4.122	7.837	1.901
GISS-E2-1-H	3.970	7.554	1.903
IPSL-CM6A-LR	4.080	7.983	1.957
MIROC6	3.242	7.780	2.400
MRI-ESM2-0	**3.416**	**7.532**	**2.205**
Ensemble Mean	*3.825 ± 0.32*	*7.738 ± 0.16*	*2.036 ± 0.19*

Table 3: The effective radiative forcing estimated using the first 20 years of global mean, annual mean data. The ensemble mean excludes CESM2 and MRI-ESM2-0.

	Relative Importance [%]			Average Fit Error [K]	
Model	**Radiative Feedback, λ**	**Heat Capacity**	**Residual**	**Abrupt-4xCO2**	**Abrupt-2xCO2**
CanESM5	94.22	6.23	-0.45	0.08	0.13
CESM2	**70.37**	**55.77**	**-26.13**	**0.09**	**0.12**
CNRM-CM6-1	135.20	-39.23	4.03	0.07	0.14
GISS-E2-1-G	471.42	-378.19	6.77	0.08	0.16
GISS-E2-1-H	98.81	3.13	-1.94	0.08	0.15
IPSL-CM6A-LR	104.52	-3.28	-1.23	0.09	0.14
MIROC6	100.20	1.13	-1.33	0.10	0.16
MRI-ESM2-0	**45.14**	**73.88**	**-19.03**	**0.07**	**0.10**
Ensemble Mean	*139.98 ± 136.46*	*-35.07 ± 143.11*	*-4.91 ±11.47*	*0.08 ± 0.01*	*0.14 ± 0.01*
Ensemble Mean (excluding GISS-E2-1-G)	*106.59 ± 16.41*	*-6.40 ± 18.67*	*-0.19 ± 2.42*		

Table 4: Relative importance of radiative feedbacks, effective heat capacity for each model. Ensemble mean with all models, ensemble mean without GISS-E2-1-G, CESM2, or MRI-ESM-0. The relation given in equation 13 is not true for GISS-E2-1-G, hence its exclusion from the mean.

	Abrupt-4xCO2 [w/m^2/K]	Abrupt-2xCO2 [w/m^2/K]	Difference
a) CESM2	**λ (Years 0 to 150)**	**λ (Years 0 to 150)**	**$\lambda_{2x} - \lambda_{4x}$**
SW CRE	0.64	0.41	**-0.23**
LW CRE	-0.14	-0.30	**-0.16**
SW Clear Sky	0.57	0.60	0.03
LW Clear Sky	-1.83	-2.00	**-0.17**

	Abrupt-4xCO2 [w/m^2/K]	Abrupt-2xCO2 [w/m^2/K]	Difference
b) MRI-ESM2-0	**λ (Years 0 to 150)**	**λ (Years 0 to 150)**	**$\lambda_{2x} - \lambda_{4x}$**
SW CRE	-0.15	-0.34	**-0.19**
LW CRE	0.03	0.07	0.04
SW Clear Sky	0.87	0.84	0.03
LW Clear Sky	-1.96	-2.02	**-0.06**

Table 5: The shortwave cloud, longwave cloud, shortwave clear sky, and longwave clear sky radiative components for CESM2 (a) and MRI-ESM2-0 (b) for abrupt-2xCO2 and abrupt-4xCO2.

	AMOC Decline (Sv)				
Model	**PiControl**	**ΔAbrupt-4xCO2**	**ΔAbrupt-2xCO2**	**½ Δ4xCO2 – Δ2xCO2**	**ΔAbrupt-2xCO2 / ΔAbrupt-4xCO2**
CanESM5	12.31	-7.95	-3.13	0.85	39%
CESM2	**23.09**	**-20.86**	**-15.05**	**-4.62**	**72%**
CNRM-CM6-1	18.08	-16.10	-10.26	-2.21	63%
GISS-E2-1G	14.88*	-12.23	+9.93	16.07	-81%
MIROC6	19.22	-15.58	-	-	-
MRI-ESM2-0	**21.50**	**-20.21**	**-19.03**	**-8.92**	**94%**

Table 6: The piControl mean AMOC strength, the change in the final 30 years of abrupt-4xCO2 and abrupt-2xCO2 from the piControl value (i.e. the AMOC decline at each forcing), the difference across forcings, and the ratio. (*) Note, GISS-E2-1-G does not have piControl streamfunction data available, so the first ten and final 30 years are used.

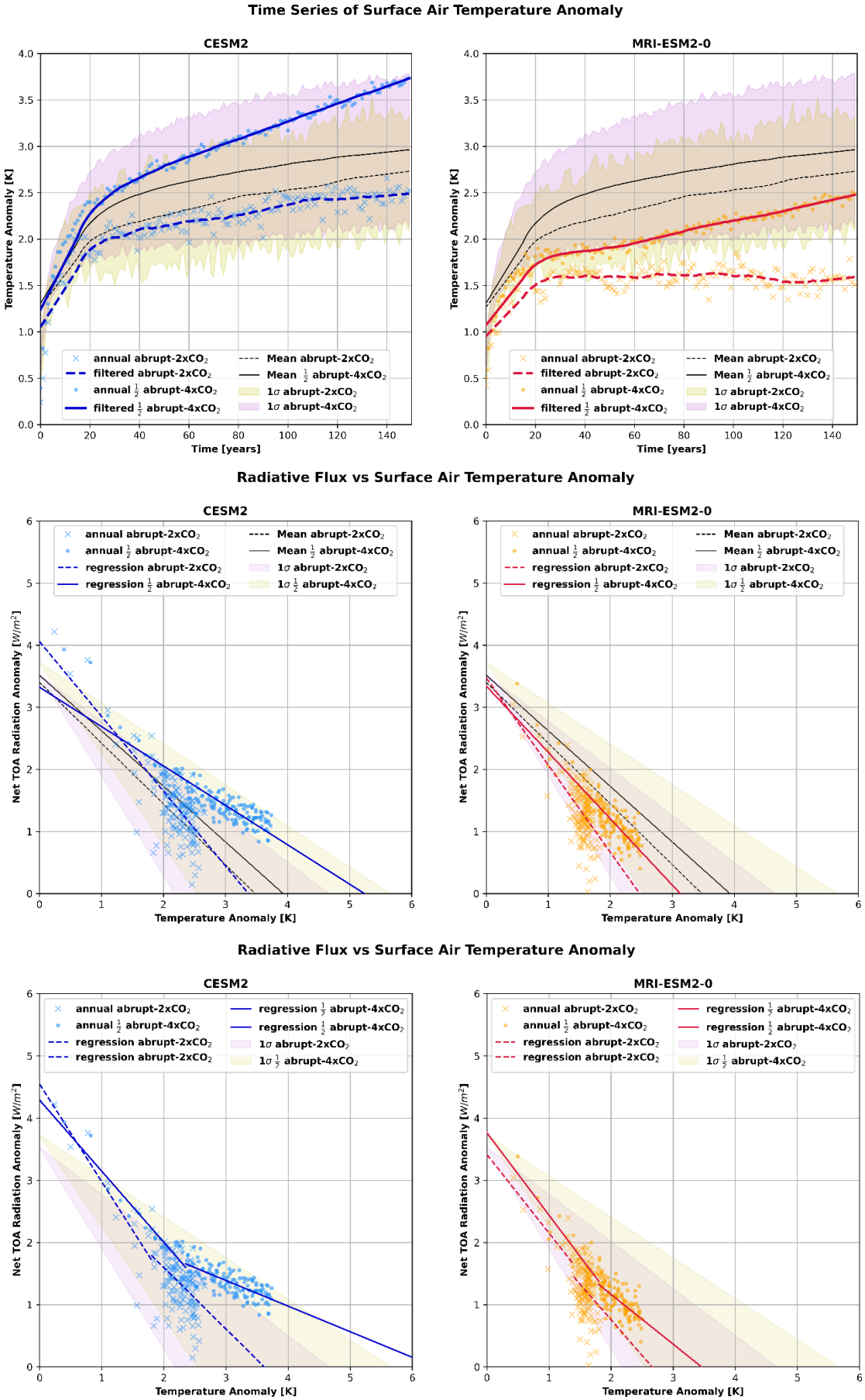

Figure 1:

(a, b) Surface temperature anomaly of CESM (blue and navy) and MRI-ESM2-0 (red and orange) and the ensemble average (excluding CESM2 and MRI-ESM2-0; black) against time. The dots are yearly averages while the lines are the yearly data with a first order polynomial filter over a 31-year window. Solid lines represent abrupt-4xCO2 and dashed lines represent abrupt-2xCO2. The filled in areas represent one standard deviation from the ensemble mean for abrupt-2xCO2 (yellow) and abrupt-4xCO2 (purple).

(c, d) The net TOA radiative flux anomaly of CESM (blue and navy) and MRI-ESM2-0 (red and orange) and the ensemble average (excluding CESM2 and MRI-ESM2-0; black) against surface temperature anomaly. The lines represent a linear regression, and the dots represent yearly averages. The filled areas represent the smallest to largest ECS model in the abrupt-4xCO2 (purple) and abrupt-2xCO2 (yellow) ensemble, excluding CESM2 and MRI-ESM2-0.

(e, f) As (c) and (d), but the regressions are over the years 0 to 20 and 21 to 150 respectively.

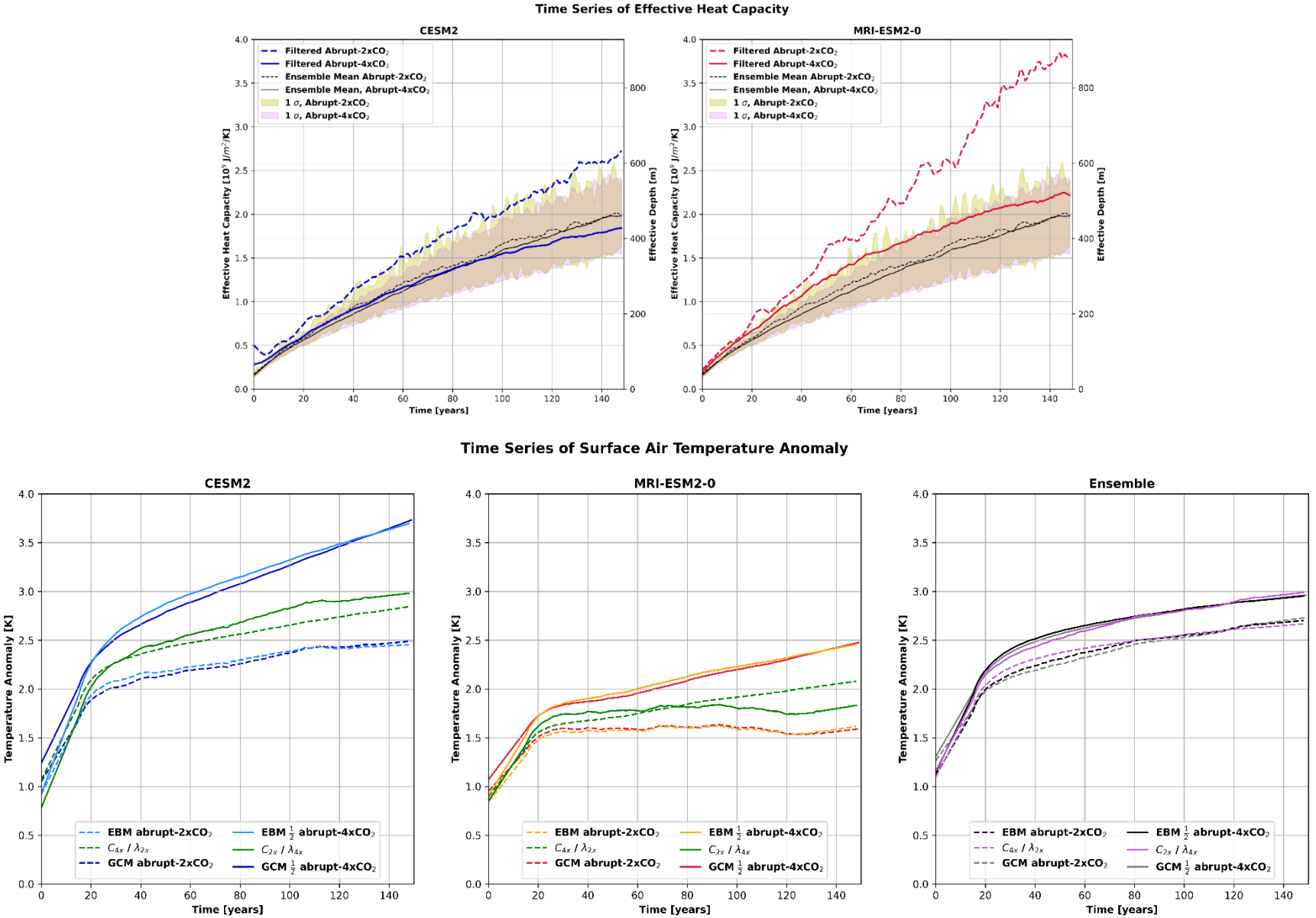

Figure 2: (a, b) The effective heat capacity of CESM2 (navy), MRI-ESM2-0 (red), and the ensemble mean (black) for abrupt-4xCO2 (solid) and abrupt-2xCO2 (dashed) over time. The filled areas represent 1 standard deviation from the mean for abrupt-4xCO2 (purple) and abrupt-2xCO2 (yellow). (c, d, e) The surface temperature anomaly of the EBM (blue, orange, black), the GCMs (navy, red, grey), and the experiments (green and purple).

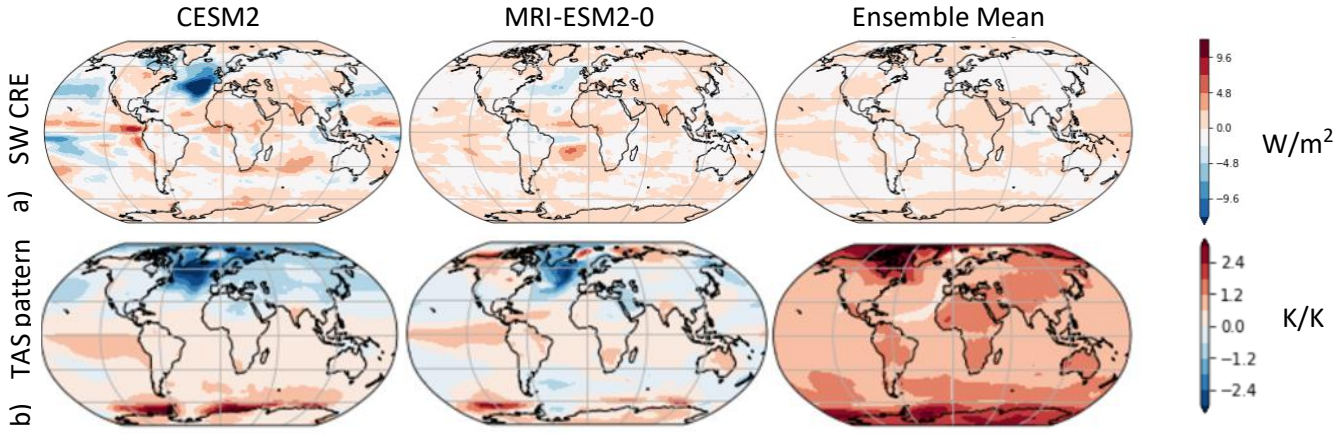

Figure 3: The APRP shortwave cloud radiative effect (a) and the temperature pattern (defined as the regression of the local temperature time series against the globally averaged temperature time series) for CESM2, MRI-ESM2-0, and the ensemble mean.

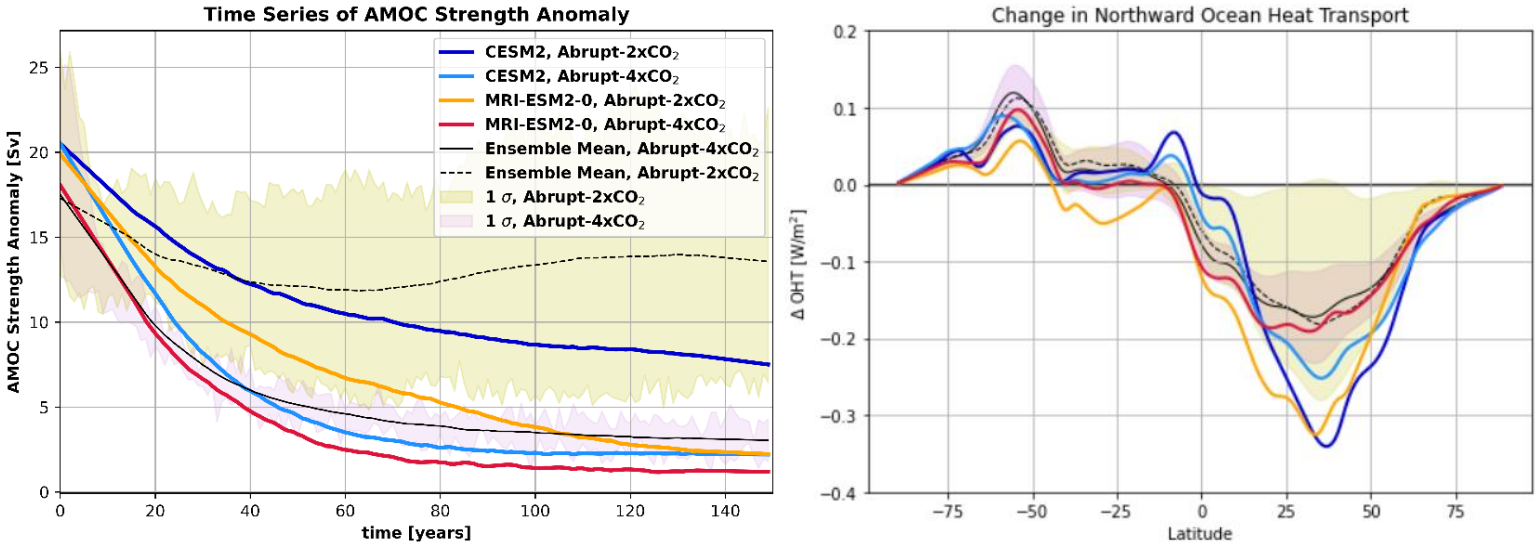

Figure 4: The AMOC strength anomaly (a) and the change in northward ocean heat transport over time (defined as the years 21 to 150 minus the years 0 to 20) (b) for CESM2 (navy and blue), MRI-ESM2-0 (orange and red), and the ensemble mean (black solid and dotted). The AMOC strength anomaly is filtered using a one-degree polynomial filter with a 5-year window. The filled in areas represent 1 standard deviation from the mean for abrupt-2xCO2 (yellow) and abrupt-4xCO2 (purple). Note, MRI-ESM2-0 is the only model for which the year one value of the AMOC strength is different across forcings, despite having the same branch point from piControl. Since we use annually averaged data, this difference may be due to rapid changes in the first year of the run.

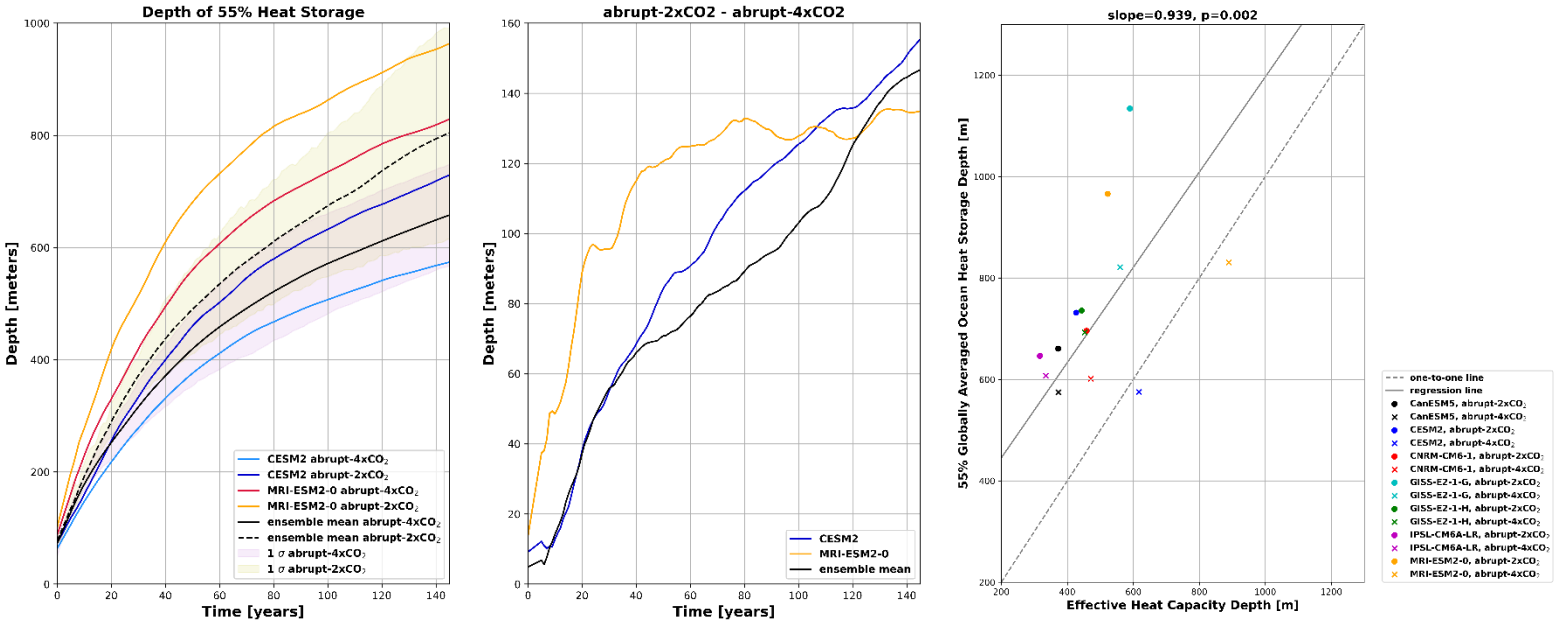

Figure 5: The globally averaged depth of 55% heat storage (a) and the difference in the depth of the 55% heat storage (b) for CESM2 (blue and navy), MRI-ESM2-0 (red and orange), and the ensemble mean (black solid and dotted). The filled in areas represent 1 standard deviation from the mean for abrupt-2xCO2 (yellow) and abrupt-4xCO2 (purple). (c) A scatter plot of the final five-year mean of the depth of 55% heat storage against the final five-year mean of the effective heat capacity depth for each model. The solid grey line represents a linear regression of those points including a point at (0, 0). The dotted grey line represents a line with a slope of 1. The linear regression has a slope of 0.939 m/m and a p value less than 0.05.

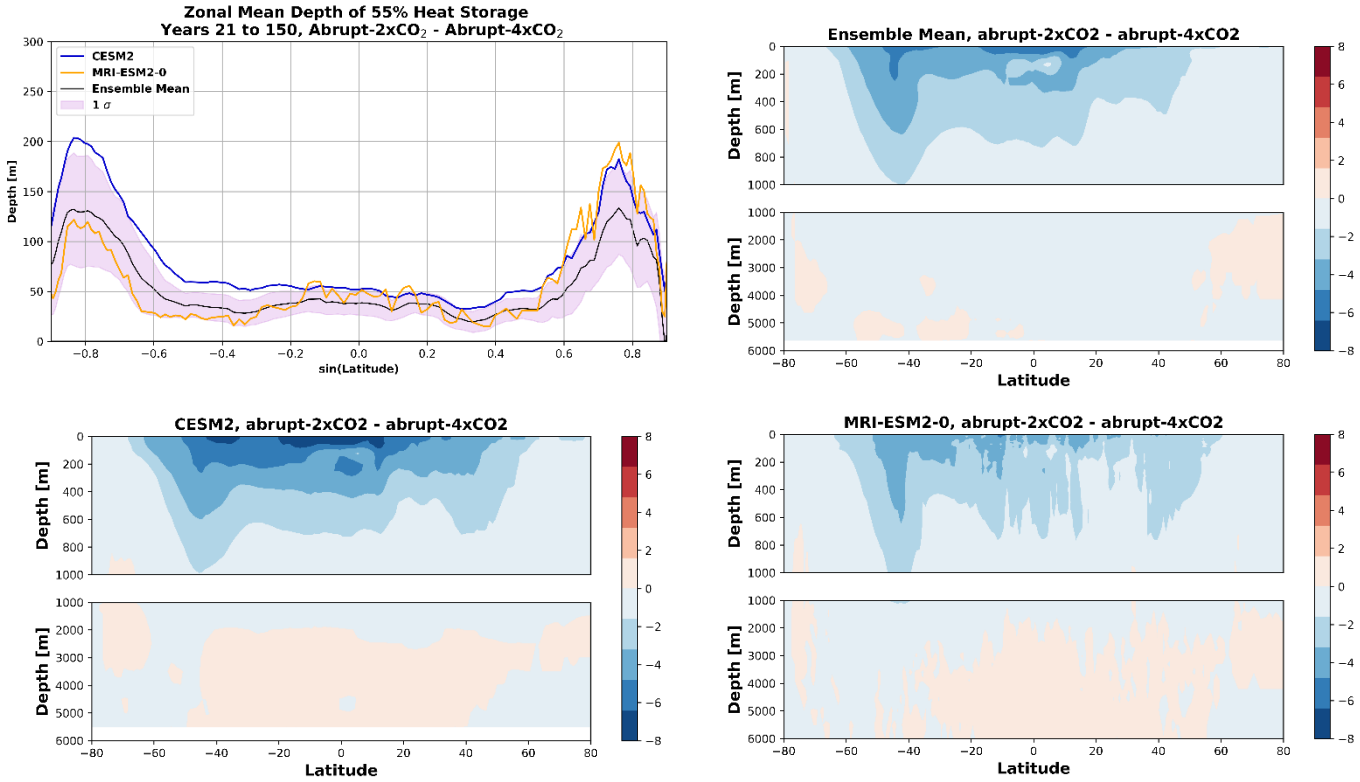

Figure 6: (a) The difference in the zonally averaged depth of 55% heat storage for CESM2 (blue), MRI-ESM2-0 (orange), and the ensemble mean (black), where the filled in area represents 1 standard deviation from the mean (purple). (b-d) The difference in zonally averaged potential temperature anomaly [K] of the ensemble mean (b), CESM2 (c), and MRI-ESM2-0 (d).

7. References

1. Ackerman, L., Danek, C., Gierz, P, et al. (2020) AMOC Recovery in a Multicentennial Scenario Using a Coupled Atmosphere-Ocean-Ice Sheet Model. GRL, 47(16). https://doi.org/10.1029/2019GL086810
2. Andrews, T., Gregory, J. M., Webb, M. J., et al. (2012) Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. GRL, 39(9). https://doi.org/10.1029/2012GL051607
3. Andrews, T., Gregory, J. M., Webb, M. J. (2015) The Dependence of Radiative Forcing and Feedback on Evolving Patterns of Surface Temperature Change in Climate Models. J Clim, 28(4), 1630-1648. https://doi.org/10.1175/JCLI-D-14-00545.1
4. IPCC (2021) Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
5. Armour, K. C. (2017) Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. Nature Clim Change, 7, 331-335. https://doi.org/10.1038/nclimate3278
6. Mett Office (2021) Mauna Loa carbon dioxide forecast for 2021. UK Met Office. Accessed 15 March 2022. https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/long-range/forecasts/co2-forecast-for-2021
7. Bonnet, R., Swingedouw, D., Gastineau, G., et al (2021) Increased risk of near term global warming due to a recent AMOC weakening. Nature Comm, 12(6108). https://doi.org/10.1038/s41467-021-26370-0
8. Byrne, B. & Goldblatt, C. (2013) Radiative forcing at high concentrations of well-mixed greenhouse gases. GRL, 41(1), 152-160. https://doi.org/10.1002/2013GL058456
9. Cheng, W., Chiang, J. C. H., Zhang, D. (2013) Atlantic Meridional Overturning Circulation (AMOC) in CMIP5 Models: RCP and Historical Simulations. J Clim, 26(18), 7187-7197. https://doi.org/10.1175/JCLI-D-12-00496.1
10. Dong, Y., Armour, K. C., Zelinka, M. D., et al. (2020) Intermodel Spread in the Pattern Effect and Its Contribution to Climate Sensitivity in CMIP5 and CMIP6 Models. J Clim, 33(18), 7755-7775. https://doi.org/10.1175/JCLI-D-19-1011.1
11. Drijfhout, S., van Oldenborgh, G. J., Cimatoribus, A. (2012) Is a Decline of AMOC Causing the Warming Hole above the North Atlantic in Observed and Modeled Warming Patterns?. J Clim, 25(24), 8373-8379. https://doi.org/10.1175/JCLI-D-12-00490.1
12. Drijfhout, S. (2015) Competition between global warming and an abrupt collapse of the AMOC in Earth's energy imbalance. Sci Rep, 5(14877). https://doi.org/10.1038/srep14877
13. Donohoe, A., Armour, K. C., Pendergrass, A. G. (2014) Shortwave and longwave radiative contributions to global warming under increasing CO2. PNAS, 111(47), 16700-16705. https://doi.org/10.1073/pnas.1412190111
14. Etminan, M., Myhre, G., Highwood, E. J., et al. (2016) Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. GRL, 43(24). https://doi.org/10.1002/2016GL071930
15. Eyring, V., Bony, S., Meehl, G. A., et al. (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci. Model Dev., 9(5) https://doi.org/10.5194/gmd-9-1937-2016

16. Geoffroy, O., Saint-Martin, D., Ribes, A. (2012) Quantifying the sources of spread in climate change experiments. GRL, (39)24. https://doi.org/10.1029/2012GL054172

17. Geoffroy, O., Saint-Martin, D., Olivie, D. J. L., et al. (2013) Transient Climate Response in a Two-Layer Energy-Balance Model, Part I. J. Clim., 26(6), 1841-1857. https://doi.org/10.1175/JCLI-D-12-00195.1

18. Geoffroy, O. & Saint-Martin, D (2020) Equilibrium- and Transient-State Dependencies of Climate Sensitivity: Are They Important for Climate Projections?. J Clim, 33(5), 1863-1879. https://doi.org/10.1175/JCLI-D-19-0248.1

19. Good, P., Lowe, J. A., Andrews, T., et al. (2015) Nonlinear regional warming with increasing CO2 concentrations. Nat. Clim Change, 5, 138–142. https://doi.org/10.1038/nclimate2498

20. Gregory, J. M., Ingram, W. J., Palmer, M. A., et al. (2004) A new method for diagnosing radiative forcing and climate sensitivity. GRL. 31(3). https://doi.org/10.1029/2003GL018747

21. Gregory, J. M., Andrews, T., Good, P. (2015) The inconstancy of the transient climate response parameter under increasing CO2. Royal Society. https://doi.org/10.1098/rsta.2014.0417

22. Hu, A., Gerald, M. A., Han, W., et al. (2015) Effects of the Bering Strait closure on AMOC and global climate under different background climates. Progress in Oceanography, 132, 174-196. https://doi.org/10.1016/j.pocean.2014.02.004

23. Hu, A., Van Roekel, L., Weijer, W., et al. (2020) Role of AMOC in Transient Climate Response to Greenhouse Gas Forcing in Two Coupled Models. J Clim, 33(14), 5845-5859. https://doi.org/10.1175/JCLI-D-19-1027.1

24. Jackson, L.C., Kahana, R., Graham, T. et al. (2015) Global and European climate impacts of a slowdown of the AMOC in a high resolution GCM. Clim Dyn 45, 3299–3316. https://doi.org/10.1007/s00382-015-2540-2

25. Jackson, L. C. & Wood, R. A. (2020) Fingerprints for Early Detection of Changes in the AMOC. J Clim, 33(16), 7027-7044. https://doi.org/10.1175/JCLI-D-20-0034.1

26. Kaspar, F., Schulzweida, U., Muller, R. (2010) "Climate data operators" as a user-friendly processing tool for CM SAF's satellite-derived climate monitoring products. EUMETSAT Meteorological Satellite Conference. https:/doi.org/10.13140/RG.2.2.20422.68165

27. Kostov, Y., Armour, K. C., Marshall, J. (2014) Impact of the Atlantic meridional overturning circulation on ocean heat storage and transient climate change. GRL, 41(6). https://doi.org/10.1002/2013GL058998

28. Leach, N. J., Jenkins, S., Nicholls, Z., et al. (2020) FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. Geosci. Model Dev., 14. https://doi.org/10.5194/gmd-14-3007-2021

29. Levitus, S., Antonov, J., and Boyer, T. (2005), Warming of the world ocean, 1955–2003, GRL, 32, L02604, doi:10.1029/2004GL021592.

30. Lin, Y.-J., Hwang, Y.-T., Ceppi, P., et. al. (2019). Uncertainty in the Evolution of Climate Feedback Traced to the Strength of the Atlantic Meridional Overturning Circulation. GRL, 46, 12,331-12,339. https://doi.org/10.1029/2019GL083084

31. Liu, M., Vecchi, G., Soden, B., et al. (2021) Enhanced hydrological cycle increases ocean heat uptake and moderates transient climate change. Nature Clim Change, 11, 848-853. https://doi.org/10.1038/s41558-021-01152-0

32. Lobelle, D., Beaulieu, C., Livina, V., et al. (2020) Detectability of an AMOC Decline in Current and Projected Climate Changes. GRL, 47(20). https://doi.org/10.1029/2020GL089974

33. Meraner, K., Mauritsen, T., Aiko, V. (2013) Robust increase in equilibrium climate sensitivity under global warming. GRL, 40(22). https://doi.org/10.1002/2013GL058118

34. Mitevski, I., Orbe, C., Chemke, R. (2021) Non-Monotonic Response of the Climate System to Abrupt CO2 Forcing. GRL, 48(6). https://doi.org/10.1029/2020GL090861

35. Rugenstein, M. A. A., Winton, M., Stouffer, R. J., et al. (2013) Northern High-Latitude Heat Budget Decomposition and Transient Warming. J Clim, 26(2), 609-621. https://doi.org/10.1175/JCLI-D-11-00695.1

36. Rugenstain, M. A. A., Bloch-Johnson, J., Abe-Ouchi, A., et al. (2019) LongRunMIP: Motivation and Design for a Large Collection of Millennial-Length AOGCM Simulations. J. Clim, 100(12), 2551-2570. https://doi.org/10.1175/BAMS-D-19-0068.1

37. Smith, C. J., Kramer, R. J., Myhre, G., et al. (2020) Effective radiative forcing and adjustments in CMIP6 models, Atmos. Chem. Phys., 20, 9591–9618, https://doi.org/10.5194/acp-20-9591-2020, 2020.

38. Taylor, K. E., Crucifix, M., Braconnot, P., et al. (2007) Estimating Shortwave Radiative Forcing and Response in Climate Models. J. Clim., 20(11), 2530-2543. https://doi.org/10.1175/JCLI4143.1

39. Trossman, D. S., Palter, J. B., Merlis, T. M., et al. (2016) Large-scale ocean circulation-cloud interactions reduce the pace of transient climate change. GLR, 43(8). https://doi.org/10.1002/2016GL067931

40. Weaver, A. J., Sedlacek, J., Eby, M., et al (2012) Stability of the Atlantic meridional overturning circulation: A model intercomparison. GRL, 39(20). https://doi.org/10.1029/2012GL053763

41. Winton, M., Griffes, S. M., Samuels, B. L, et al. (2013) Connecting Changing Ocean Circulation with Changing Climate. J. Clim, 26. https://doi.org/10.1175/JCLI-D-12-00296.1

42. Zelinka, M. D., Myers, T. A., McCoy, D. T., et al. (2020) Causes of Higher Climate Sensitivity in CMIP6 Models. GRL, 47(1). https://doi.org/10.1029/2019GL085782

43. Zhang, R., Kang, S. M., Held, I. M. (2010) Sensitivity of Climate Change Induced by the Weakening of the Atlantic Meridional Overturning Circulation to Cloud Feedback. J Clim, 23(2), 378-389. https://doi.org/10.1175/2009JCLI3118.1