Climate model code genealogy and its relation to climate feedbacks and sensitivity

Peter Kuma[1], Frida A.-M. Bender[1], and Aiden R. Jönsson[1]

[1]Department of Meteorology (MISU), Stockholm University, Stockholm, Sweden

Correspondence: Peter Kuma (peter.kuma@misu.su.se)

Abstract. Contemporary general circulation (GCM) models and Earth system models (ESMs) are developed by a large group of modelling centres internationally. They use a broad range of implementations of climate dynamics and physical parametrisations, allowing for structural (code) uncertainty to be partially quantified with multi-model ensembles (MMEs). However, many models in the MMEs of the Climate Model Intercomparison Project (CMIP) have a common development history due to the widespread practice of sharing code and parametrisations within and between modelling centres. This makes results from different models statistically dependent, potentially introducing biases in MME statistics. This situation became more pronounced in CMIP6 compared to CMIP5 due to the proliferation of model runs contributed by the same model, and due to the fact that several models predict much higher effective climate sensitivity (ECS) than multiple evidence assessments such as the Intergovernmental Panel on Climate Change Sixth Assessment Report, and this means that some MME statistics differ from multiple evidence estimates. Previous research investigating effects of model inter-dependence has focused on model output and code dependence, but model code genealogy of CMIP models has not been fully analysed. We present a full reconstruction of CMIP3, CMIP5 and CMIP6 code genealogy of 167 atmospheric models, GCMs and ESMs (of which 114 participated in one of the CMIP phases) based on available literature and online resources, with a focus on inheritance in the atmospheric component and atmospheric physical parametrisations. We developed a model code weighting method based on the model code genealogy for the purpose of analysing the impact of such weighting on MME means. We assess the implications of such weighting on ECS, climate feedbacks, forcing and global mean near-surface air temperature, as well as simpler weighting methods based on model family, institute and country in CMIP5 and CMIP6. In some cases the impact is found to be substantial and can partially reconcile the differences in MME means between CMIP5 and CMIP6. We show that some model families have a propensity to be relatively warm or cold in the CMIP5 and CMIP6 experiments. Our method is complementary to the existing methods based on model output clustering. The presented results can help in understanding structural dependencies between CMIP models, and the proposed code and family weighting methods can be used in MME assessments to ameliorate model structure sampling biases.

1 Introduction

General circulation models (GCMs) and Earth system models (ESMs) are currently the most sophisticated tools for studying paleontological, historical, present-day and future climate. The development of GCMs has a long history, interlinked with the

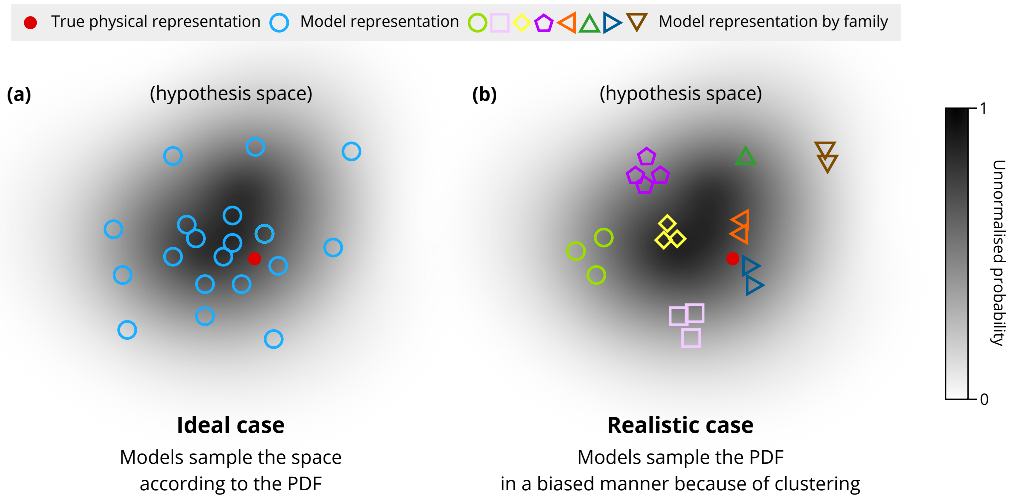

Figure 1. A theoretical illustrative example of model sampling of the model hypothesis space (model structural uncertainty), representing realisations of physical climate processes (model structure). The shading indicates a probability density function (PDF) quantifying our collective belief that a certain representation is true. In an ideal case **(a)**, models are unbiased samples from this PDF, allowing us to estimate the PDF from a multi-model ensemble (MME). In reality **(b)**, they form clusters because of structural model dependence (code sharing) as assumed and discussed in the introduction, sampling the PDF in a biased manner. Weighted sampling is necessary to estimate the PDF from such an MME. The unknown true physical representation, not coinciding with the PDF maximum or mean, is indicated by a red dot. For illustrative purposes only, the hypothesis space is visualised in a 2-dimensional space. Model marker colours (shapes) in **(b)** indicate different hypothetical model families, within which models are structurally related. Note that the PDF represents model structure and might not correlate with model output PDF.

development of numerical weather prediction (NWP) models (Lynch, 2008). Intercomparison of climate models dates back to the late 1980s when the Atmospheric Model Intercomparison Project (AMIP) started comparing atmospheric models under standardised conditions and model output (Touzé-Peiffer et al., 2020). This was followed by the Climate Model Intercomparison Project (CMIP) phase 1 and 2 in 1996 and 1997, respectively, and these informed the Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change (IPCC). CMIP3 (Meehl et al., 2007) was the first time that model output became openly available to all researchers, and therefore enabled a wide research of climate models together as multi-model ensembles (MMEs). However, this came with difficulties because such a multi-model dataset was not designed to represent structural model uncertainty in an unbiased way (Abramowitz et al., 2019). The two most recent CMIP phases are phase 5 (Taylor et al., 2012) and phase 6 (Eyring et al., 2016, 2019).

Modern climate models such as GCMs and ESMs are highly complex software, consisting of many components, modules and configuration parameters. Usually, components such as the atmosphere, ocean, land, sea ice, chemistry, biology and others are coupled together continuously during a simulation (Alexander and Easterbrook, 2015). These components may be divided into subcomponents, modules or schemes representing various physical parametrisations, such as the radiative transfer in the atmospheric component. Components and subcomponents sometimes can be easily replaced with others, or they can be turned on or off depending on the configuration. These model parts have been shared relatively freely between different models in the same modelling group as well as between groups internationally. Alexander and Easterbrook (2015) analysed software components of models by directly analysing their source code, showing significant sharing of components between models thanks to their highly modular nature. Furthermore, parametrisations documented in literature were implemented in a variety of models, meaning that they use many of the same parametrisations for certain physical processes. This development approach leads to structural model dependence, which could mean that their model output is more similar than what would be expected from structurally independent models. Understanding of model structural dependence is further complicated by the fact that only few models have publicly available source code. The practice of 'forking' code, when a new branch of a code base is created under a new name, is very common in software development. This is also the case with climate models, where different modelling groups base their work on forking of an existing model from the same or a different modelling group. This process can be quite opaque to the end-users, who might, without access to further context, assume that a different model name implies that the model is entirely independent. We can expect that model code bases which are open source (such as the Community Earth System Model (CESM)) or licensed widely within international consortia (such as the Integrated Forecasting System (IFS)/ARPEGE and Hadley Centre Global Environmental Model (HadGEM)) are more highly represented in model ensembles due to the ease of sharing code (Sanderson et al., 2015b). This is potentially in contrast to proliferation of code which produces the best results, which could otherwise arise if all model code were openly available. As discussed below, what constitutes 'best results' may be difficult to quantify and is not guaranteed to coincide with best projections. Guilyardi et al. (2013) initiated better model and experiment metadata collection within CMIP5 to provide pertinent information to those performing research based on model comparisons.

Because all models are imperfect representations of reality, they are affected by various uncertainties in the model output, which can be broadly categorised as data, parameter and structural uncertainty (Remmers et al., 2020). While data and parameter uncertainty can be relatively easily quantified and sampled, structural uncertainty pertaining to model code is hard to quantify or sample, and some authors noted that structural uncertainty is insufficiently sampled in CMIP MMEs (Knutti et al., 2010). Models participating in CMIP are dependent in a number of ways, including being essentially the same model with a different configuration, sharing parts of their codes, model components and schemes, using the same datasets for validation, and implementing similar parametrisations. Some authors have therefore called this MME an 'ensemble of opportunity' (Masson and Knutti, 2011; Knutti et al., 2013; Sanderson et al., 2015a; Boé, 2018), since the inclusion is based on the intent of a modelling centre to participate rather than any objective selection criteria. If model dependence is not taken into account, the calculation of means and variance and uncertainty can be biased, as well as spurious correlations (such as in emergent constraints) can arise in an MME (Caldwell et al., 2014; Sanderson et al., 2021). Remmers et al. (2020) posed a question whether model code genealogy can be inferred from model output. Using a modular modelling framework, they generated a model ensemble of hydrological models by sampling the model 'hypothesis space' and compared its genealogies based on model code and model output. They found that it was not possible to infer complete model code genealogy based on model output because performance of the inference was low. It is possible that the same would partially apply to much more complex models like GCMs and ESMs, and model code relationship needs to be studied in order to sample the model hypothesis space. Pennell and

Reichler (2011) tried to quantify the effective number of models in an MME of 24 CMIP3 models based on model output error similarity, and found this to be about 8, with diminishing returns with increasing number of models. Sanderson et al. (2015b) reached a similar conclusion, and found that the number of independent models (based on model output) in CMIP5 is much smaller than the total.

The simplest approach to analysing an MME is 'model democracy', where each model is given an equal weight in statistical calculations. More sophisticated approaches proposed to address model dependence include weighting or selecting models. Selecting models can be regarded as an extreme form of weighting. Often suggested weighting methods are based on model performance ('model meritocracy'), model output or code dependence and diversity. The topic of climate model dependence and genealogy has been covered in many previous studies, most of which used the dependence of the model output (Jun et al., 2008a, b; Masson and Knutti, 2011; Knutti et al., 2013; Bishop and Abramowitz, 2013; Sanderson et al., 2015a; Haughton et al., 2015; Mendlik and Gobiet, 2016), while a focus on code dependence has been relatively more rare (Alexander and Easterbrook, 2015; Steinschneider et al., 2015). Boé (2018) distinguishes these two approaches as 'a posteriori' and 'a priori'. Knutti et al. (2013) developed a CMIP5 model genealogy based on hierarchical clustering of model output. A more simple approach is 'institutional democracy', where one model per modelling centre is selected, and 'component democracy', where models are selected to represent different model components (Abramowitz et al., 2019). Edwards (2000a, b, 2011) constructed a partial 'family tree' of atmospheric GCMs based on their code heritage. Boé (2018) summarised a modelling group, atmospheric, oceanic, land and sea ice components of CMIP5 models and how they relate to proximity of the model results. However, the code dependence of all CMIP3, CMIP5 and CMIP6 models has not been analysed. Partially, such understanding is limited by the availability of the source code. This contributes to the treatment of models as 'black boxes' by the research community. Haughton et al. (2015) compared simple weighting with model performance and model output dependence weighting. They found performance weighting improved mean (as expected) but degraded variance estimation, and dependence weighting improved both. Steinschneider et al. (2015) identified close correlations between model output of models of the same family even on a regional scale, and showed that clustering of similar models can result in narrowing the MME variance attributable to intermodel correlations.

Reducing the size of an MME to a set of independent models is a relatively simple method of avoiding model dependence. Sanderson et al. (2015b) noted that if only one model per institute is permitted in an MME, it could lead to unfairly dismissing models which are substantially different, and overestimating independence in cases where code is shared between institutes. Weighting models by country can have some merit due to the fact that models are sometimes developed with a focus on accuracy over the region where the institute is located, and a model might be more extensively validated against data from observations in the region. For example, the New Zealand Earth System Model (NZESM) (in practice sharing common development with HadGEM/UKESM) was developed to reduce Southern Ocean biases (Williams et al., 2016), the Indian Institute of Tropical Meteorology ESM (IITM ESM) had a special focus on the South Asian monsoon (Krishnan et al., 2021), the Australian Community Climate and Earth System Simulator coupled model (ACCESS-CM) has a focus on reducing uncertainties over the Australian region (Bi et al., 2013), and the Energy Exascale Earth System Model (E3SM) aimed to support the U.S. energy sector decisions (Golaz et al., 2019). Weighting models by errors relative to observations is complicated by the fact that there can be a decoupling between a climate model's accuracy in representing present-day and historical climate variables and its accuracy in representing the change (or trend) of the variables under a climate scenario (Jun et al., 2008a; Zelinka, 2022). Thus, performance of a model in future climate projections cannot be fully inferred from its performance in present-day and historical climate. Performance weighting can also favour models which are better tuned to present-day, historical or paleontological observations by compensating biases. It is possible that model quality cannot be estimated solely from the model output due to the fact that some models might represent physics more consistently with our knowledge of fundamental physics, yet give inferior output when compared to observations if they have fewer compensating biases or were tuned less to represent present-day or historical observations. Apart from explicit model weighting or selection choices, seldomly recognised implicit choices based on values (other than openness, objectivity, evidence and impartiality) influence model development, evaluation, selection, weighting, interpretation and communication of results (Pulkkinen et al., 2022a, b; Lenhard and Winsberg, 2010; Winsberg, 2012).

We can define the structure (code) of a model as based on a set of hypotheses about reality as well as computational realisations of such hypotheses. A desirable feature of an MME would be that models represent samples from the hypothesis space with probability equal to our degree of belief that the hypothesis is true (note that this is different from a uniform sampling of the hypothesis space, which would be both impossible and undesirable due to its size). However, this is rarely the case with existing MMEs, nor it is easily quantifiable. It is generally not desirable that the model output of individual models in an MME is the most unique, because one would still want all models to converge as closely as possible on the true representation of physical processes. Models can be close in their model output because they are convergent on the best representation of reality or because of code similarity, and this limits the use of model output as a measure of model dependence.

As a conceptual model (Fig. 1), we can consider models in an MME to be samples corresponding to representations of a physical reality in a hypothesis space. Here, representation is supposed to mean code which produces output for given initial and boundary conditions, i.e. without considering internal variability. While the true physical representation is unknown and impossible to simulate due to computational constraints, our collective belief that a given representation is true can be conceptualised theoretically by a probability density function (PDF). Ideally, models in an MME are independent samples from this PDF (Fig. 1a). In actual MMEs (Fig. 1b), however, models are dependent and tend to be clustered together for reasons incompatible with the PDF, such as inclusion of several configurations or resolutions of a single model, selective sharing of code between models for reasons other than meritocracy (such as availability or political and organisational decisions) or model output availability. Therefore if a PDF or its statistics are estimated from this MME, they will be biased compared to the actual PDF. The aim is then to compensate for this bias with appropriate model weighting, selection or more sophisticated techniques such as emergent constraints. Even if we could estimate the PDF in an unbiased way, its highest likelihood point or mean is unlikely to coincide with the true physical representation, because such a PDF only represents our belief that a given physical representation is true (limited by our knowledge). Note that model dependence itself does not preclude that an estimate of the PDF is unbiased. For example, in the Metropolis algorithm (Metropolis et al., 1953), an unbiased estimate of a PDF is generated by sequentially producing a chain of samples which are close to each other. After a large enough number of iterations, an unbiased estimate of the PDF can be inferred from the collection of all samples, despite close correlation between adjacent samples in the chain.

None of the model weighting methods mentioned above are without issues. Performance weighting can disregard models whose physics representation is relatively far from the most likely representation but still plausible, thus artificially narrowing the spread. Model dependence weighting based on output or code can disregard models which are close to another model but were chosen to be based on this model because of its perceived quality, thus preventing such an MME from narrowing down on the true representation of climate physics. Dependency weighting based on output can mistakenly identify two models as similar when they are in fact independent, or fail to identify models with significant code dependence. Weighting based on diversity can give too much weight to outliers and too little weight on models more densely clustered around the most likely representation, thus artificially increasing the spread.

Recently, multiple models participating in CMIP6 (Eyring et al., 2016) predicted much higher effective climate sensitivity (ECS) than the assessed range of the IPCC Sixth Assessment Report (AR6). This was exacerbated by the fact that some models contributed multiple runs, making simple multi-model means potentially unreliable. Voosen (2022) cautioned that using models which predict too much warming compared to the range assessed by the AR6 can produce wrong results, and therefore model democracy should be replaced with model meritocracy. Partly due to the limitations of the simple multi-model mean, the authors of the AR6 departed from the use of multi-model means to quantify ECS and transient climate response (TCR) and instead used a multi-evidence approach similar to Sherwood et al. (2020), although a simple multi-model mean is used in other parts of the report.

2 Motivation and objectives

Code dependence in CMIP models is not well explored, especially when in comes to code sharing between modelling centres. This hinders model evaluation studies, which sometimes regard the CMIP MME as an opaque set of models (e.g. Meehl et al. (2020); Schlund et al. (2020); Zelinka et al. (2020), but also many parts of AR6). To gain understanding of the whole MME, we map the code genealogy of all CMIP atmosphere GCMs (AGCMs), atmosphere–ocean GCMs (AOGCMs) and ESMs. Much of the information about code dependence is already available in literature as well as CMIP model metadata and online resources of modelling centres, but it has not been systematically organised across CMIP phases. When determining the code relations, our focus is on the atmospheric component and atmospheric physics due to the fact that currently they are the main source of model uncertainty when in comes to climate sensitivity and cloud feedback due to uncertainties in cloud simulation, and the spread in model ECS is currently dominated by the spread in the cloud feedback (Wang et al., 2021; Forster et al., 2021; Zelinka et al., 2020). Steinschneider et al. (2015) also identified the atmospheric component as being a particularly important factor in the similarity of climate projections of temperature and precipitation between models. However, other model components such the ocean component can also have an impact on the feedbacks and climate sensitivity (Gjermundsen et al., 2021). We present a model weighting algorithm based on the model code genealogy, and investigate if it makes a difference in multi-model means of ECS, effective radiative forcing (ERF), climate feedbacks and global mean near-surface temperature (GMST) time series.

The algorithm can be used to produce weights for any given subset of CMIP models. In addition, we explore more simple weighting methods based on model family, institute and country, and analyse whether model families differ substantially in their predictions from other model families and a simple multi-model mean.

3 Methods

3.1 Data

In our analysis we focus on AGCMs, AOGCMs and ESMs in the last three phases of CMIP (3, 5 and 6). The CMIP5 and CMIP6 model output data from the control (*piControl*), *historical*, Shared Socioeconomic Pathway 2-4.5 (*ssp245*), Representative Concentration Pathway 4.5 (*rcp45*), abrupt quadrupling of CO_2 (*abrupt-4xCO2*) and 1% yr^{-1} CO_2 increase (*1pctCO2*) experiments were acquired from the public archives on the Earth System Grid (CMIP5, 2022; CMIP6, 2022) (equivalent data from CMIP3 were not analysed here). In order to analyse model code genealogy, we performed a broad literature survey, complemented by CMIP model metadata and information available online, particularly on websites of modelling centres. In total, we traced the genealogy of 167 models, of which 114 were participating in CMIP, and the rest were related to the CMIP models and necessary for reconstructing the genealogy. The model genealogy information, including related references, is also available in the spreadsheet `models.csv` in the supplementary code, and a related list of references is provided in the supplement. Along with relations between models, we identified model institute (model development centre), the country where the institute resides and model family, defined by the oldest ancestral model in the genealogy. Model parameters such as ECS, TCR, effective radiative forcing (ERF) and climate feedbacks were sourced from Zelinka et al. (2020) and the AR6. We use effective climate sensitivity calculated by Zelinka (2022), as an approximation of equilibrium climate sensitivity.

3.2 Weighting methods

We applied several statistical weighting methods on the CMIP MMEs:

- *Simple weighting*. Every model run is given equal weight. By ‘model run’ we mean a model variant, resolution or configuration (as listed in the spreadsheet `models.csv` in the columns *CMIP3/5/6 names* in the supplementary code), not multiple simulations performed with the same model but different initial conditions.

- *Family weighting*. Model families, defined as a complete branch as shown in Fig. 2 (discussed later in Sect. 4.1), were given equal weight. This weight was further subdivided equally between models within the family.

- *Institution weighting*. Model institutes, as shown in Fig. 2 as labels on gray areas, were given equal weight. This weight was further subdivided equally between models within the institute.

- *Country weighting*. Model host countries, as shown in Fig. 2 as labels on gray areas, were given equal weight. This weight was further subdivided equally between models of the same country.

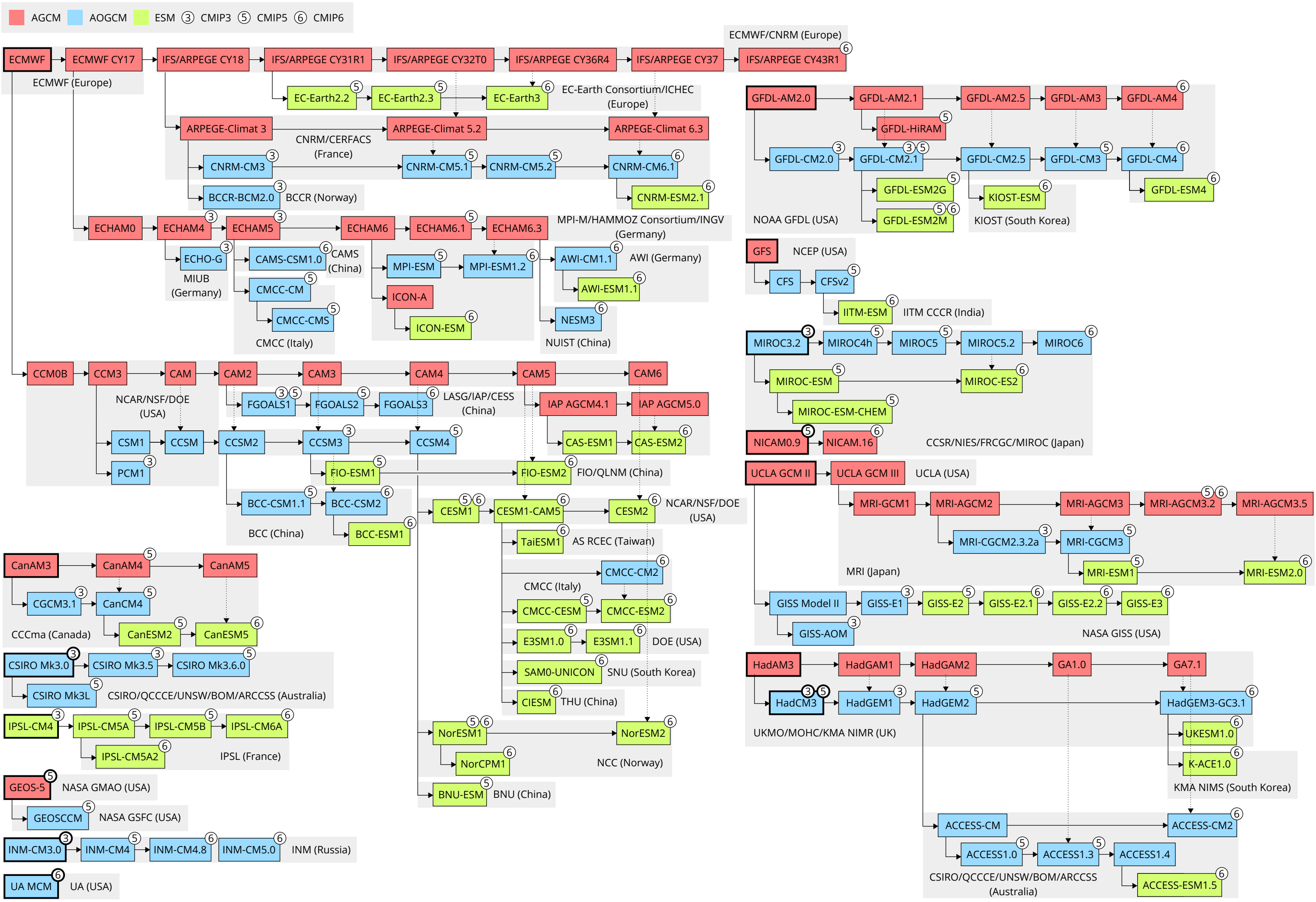

Figure 2. Model code genealogy of models participating in the Climate Model Intercomparison Project (CMIP) phase 3, 5 and 6, including their related common ancestor models. Models are distinguished by their complexity into atmosphere general circulation models (AGCMs), atmosphere–ocean GCMs (AOGCMs) and Earth system models (ESMs), indicated by colour. Horizontal arrows indicate inheritance between multiple versions of the same model. Vertical solid arrows indicate inheritance between different models. Vertical dotted arrows indicate inheritance from an AGCM to an AOGCM or ESM (this can also mean that the model is used as a component of the more complex model). The shaded boxes indicate an institute and the main country or region where the development was conducted. Numbers in circles indicate the CMIP phase. Model boxes with a thick outline indicate the oldest model of the model family. The genealogy only traces models necessary for placing the CMIP models in the graph and omits versions not included in CMIP. The genealogy was reconstructed based on available literature, CMIP metadata and online resources.

- *Code weighting*. The oldest ancestor models (marked with a thick outline in Fig. 2) were given equal weight. This weight was subdivided gradually through branches to descendant models. This method is described in detail in Appendix A.
- *Model weighting*. All models are given the same weight. This is different from the *simple weighting* – see the note below.

Note that in all of the above, if a model supplied multiple runs of different configuration or resolution, the model weight was further subdivided equally between the runs. For clarity, in the following text references to the weighting methods and weighted means corresponding to the methods above are *italicised*.

3.3 Statistical significance

Statistical significance in climate feedbacks, sensitivity and forcing in Sect. 4.3 was calculated using a Bayesian simulation with PyMC3 (Salvatier et al., 2016). The difference between a *simple* mean of models within a family and a *simple* multi-model mean was marked as significant if the magnitude difference between the two means was larger than zero with 95% probability. The PyMC3 model is provided in the supplementary code.

4 Results

4.1 Model code genealogy and model families

We traced climate model code genealogy based on available literature, with a focus on CMIP3, CMIP5 and CMIP6 models. Figure 2 presents a graph of models which includes all available CMIP AOGCMs and ESMs, except for some model subderivatives and configurations, which are grouped under a common model name. The model relations were identified with a primary focus on the atmospheric component, and in particular atmospheric physics, which is a compromise due to the fact that some models inherit multiple components (atmosphere, ocean, cryosphere, chemistry, etc.), or in some instances provide their own implementation of atmospheric dynamics while inheriting atmospheric physics from a parent model. Some models comprised multiple model runs in CMIP (configurations, resolutions or variations of components), and we grouped these thogether under a single model name. We identified 14 different model families – groups of models which share the same oldest ancestor model (marked with a thick outline in Fig. 2 and also listed in Table S1). The models come from 38 different institutes or institute groups and 15 different countries. Institutes are based on the *institute* attribute of the CMIP datasets (CMIP3, 2022; CMIP5, 2022; CMIP6, 2022) for CMIP models and reference publications or online resources for other models, separated by a slash if multiple institues were involved. Country is the country of the main institute (defined loosely as the institute credited for most of the models in the group, or where the development originated), with the exception of the European community (EC)-Earth Consortium models, for which the assumed 'country' is Europe. We recognise two kinds of model relations: a parent–child relation, when the child model is a code-derivative of the parent model with a different name (in the sense of fully or partially inheriting the code of the atmospheric component), and a relation between versions of the same model. In Fig. 2, the former is represented by a descending arrow, and the latter is represented by a horizontal arrow. Model counts per model family, country and institute in each CMIP phase are listed in Table S1.

We take an exception to the rule that a model family is defined by the oldest ancestral model for the ECMWF- and CCM-derived models, for which the model ECMWF is a common ancestor. We split this model family into two model families of ECMWF and CCM (beginning with CCM0B). This is a subjective choice made for our analysis in order to account for the fact that this split happened in very early stages of the development in the 1980s (Edwards, 2011), and the separate CCM and ECMWF model families are very large and diverse. The model families used further in our analysis are: ECMWF, CCM, CanAM, CSIRO, IPSL, GEOS, INM, UA MCM, GFDL, GFS, MIROC, NICAM, UCLA GCM and HadAM.

Some of the identified model families are relatively small, such as CSIRO, GEOS, GFS, INM, UA MCM, NICAM, with fewer than four models in CMIP, while others are very large, e.g. CCM with 28 models and ECMWF with 23 models in CMIP (here by ‘model’ we mean the main model as in Fig. 2 rather than model runs in CMIP). In terms of model runs, CCM, ECMWF and HadAM are particularly represented in CMIP6 with 32, 27 and 12 model runs, amounting to about 70% of the entire CMIP6 MME (Table S1). This means that there is a very uneven model representation in CMIP6. The situation was getting more pronounced with successive CMIP phases: in CMIP5 and CMIP3 the share of the three most represented model families in terms of model runs is smaller at 52% and 50%, respectively. The size of model families and the diversity of models within a family are clearly influenced by the availability of model code. For example, the IFS/ARPEGE model is widely licensed to participating modelling centres in Europe, and therefore is used as a basis for a multitude of different models on the continent. The CCM-derived models have publicly available source code, which has been used extensively by many different modelling centres internationally. Other models with private code are used much more narrowly, such as CanAM, CSIRO, IPSL or INM are only used by their own modelling centre (and possibly a few collaborating organisations). Publicly available or widely licensed models usually have much greater participation in CMIP and an outsized impact in the MMEs.

Relations between model code often can be complex, ranging from a model component shared with an ‘upstream’ project (such as models in the CCM family using the Community Atmosphere Model [CAM]) to models taking atmospheric physics implementation from a parent model and developing their own atmospheric dynamics. Likewise, the ocean, land, sea ice and biochemistry components are in some derived models swapped for other components. This complicates the notion of a model derivative. Because climate feedbacks in the atmosphere are currently the largest source of uncertainty in determining climate sensitivity, it is perhaps the most important model component to use as a determinant in model code genealogy. This is a subjective choice and other choices would be possible when constructing a model code genealogy.

4.2 Climate feedbacks and sensitivity

Here, we evaluate how the proposed *code weighting* and several simpler types of weighting impact climate feedbacks and climate sensitivity calculation in the CMIP MMEs. Zelinka et al. (2020) analysed climate feedbacks, ECS and ERF in CMIP5 and CMIP6. We perform the same analysis using their estimated values of model quantities (Zelinka, 2022), but with different methods of weighting. Figure 3 shows results analogous to Fig. 1 in Zelinka et al. (2020), but as means calculated using the different weighting methods relative to the *simple* multi-model mean. Following Zelinka et al. (2020), the ‘net (feedback) refers to the net radiative feedback computed directly from TOA fluxes, and the residual is the difference between the directly calculated net feedback and that estimated by summing kernel-derived components.’ The differences in feedbacks between the

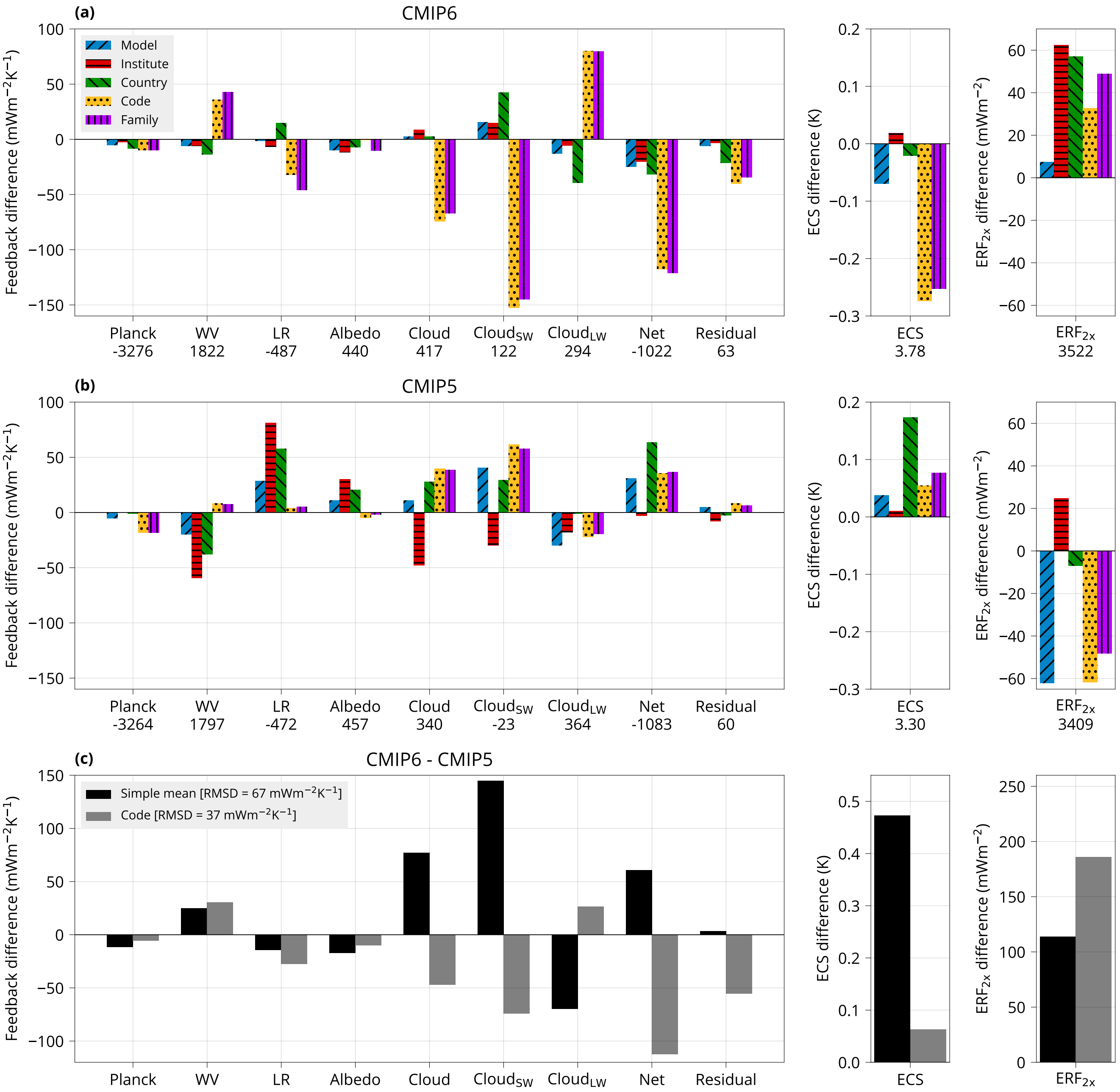

Figure 3. Climate feedbacks, effective climate sensitivity (ECS), effective radiative forcing (ERF_{2x}) in the Climate Model Intercomparison Project (CMIP) phase 5 **(a)** and 6 **(b)** under different weighting methods (*model*, *institute*, *country*, *code* and *family*) relative to a *simple* mean (Sect. 3.2). **(c)** The difference between CMIP6 and CMIP5. The legend in **(c)** shows root mean square difference (RMSD) between CMIP6 and CMIP5 (Sect. 4.2). Climate feedbacks: Planck, water vapour (WV), lapse rate (LR); surface albedo (Albedo); total cloud feedback (Cloud); shortwave cloud feedback ($\mathrm{Cloud}_{\mathrm{SW}}$); longwave cloud feedback ($\mathrm{Cloud}_{\mathrm{LW}}$); net feedback (Net); residual feedback (Residual). The underlying data are from Zelinka (2022), described in Zelinka et al. (2020).

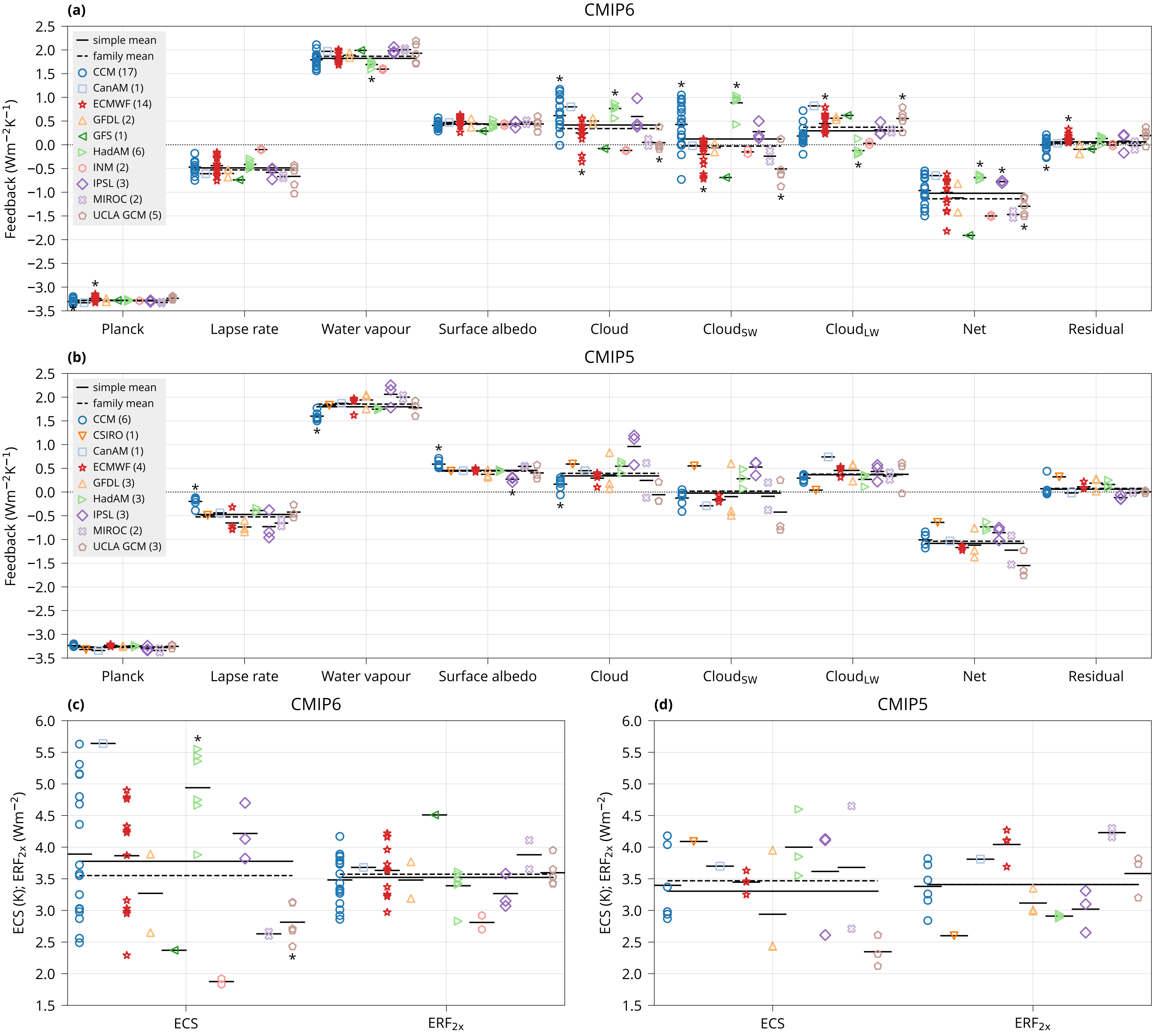

Figure 4. Climate feedbacks, effective climate sensitivity (ECS) and effective radiative forcing (ERF_{2x}) arranged by model family in the Climate Model Intercomparison Project (CMIP) phase 5 **(b, d)** and 6 **(a, c)**. Model family is identified by the oldest ancestor model. In the legend, numbers in parentheses are the number of models in the family present in the plot. Model families whose *simple* mean is significantly different (with 95% confidence) from the *simple* multi-model mean are marked with an asterisk (‘*’). The underlying data are from Zelinka (2022), described in Zelinka et al. (2020).

simple mean and the other types of weighting is up to about 150 $\mathrm{mWm^{-2}K^{-1}}$ in magnitude in CMIP6 and 80 $\mathrm{mWm^{-2}K^{-1}}$ in CMIP5. The different types of weighting often do not agree, except for the *family* and *code weighting*, which give very similar results. If we focus on the weighting methods which we expect to be the most accurate in terms of accounting for model code sharing, the *code* and *family weighting*, the largest difference from the *simple* mean is in the cloud feedbacks (total, shortwave and longwave), with relatively large difference in ECS and ERF. This is perhaps not surprising due to the very large model spread in the cloud feedbacks in the CMIP MMEs.

Interestingly, when we quantify the difference in feedback strength between the CMIP6 and CMIP5 MMEs (Fig. 3c), we see that the *code weighting* reduces the difference in cloud feedbacks between the two CMIP phases substantially. For the total cloud feedback, the magnitude difference is reduced from 77 to -47 $\mathrm{mWm^{-2}K^{-1}}$, for the shortwave (SW) cloud feedback from 145 to -74 $\mathrm{mWm^{-2}K^{-1}}$, and for the longwave (LW) cloud feedback from -70 to 27 $\mathrm{mWm^{-2}K^{-1}}$. However, the net and residual feedback magnitude difference is increased from 61 to -113 $\mathrm{mWm^{-2}K^{-1}}$ and from 3 to -55 $\mathrm{mWm^{-2}K^{-1}}$, respectively. We define root mean square difference (RMSD) between CMIP6 and CMIP5 calculated across the elementary feedbacks (Planck, water vapour (WV), lapse rate (LR), albedo, SW cloud, LW cloud) as:

$$\mathrm{RMSD} = \left(\frac{1}{n} \sum_{i=1}^{n} (\lambda_{i,\text{CMIP6}} - \lambda_{i,\text{CMIP5}})^2 \right)^{1/2}, \quad n = 6, \quad \lambda_i = (\lambda_{\text{Planck}}, \lambda_{\text{WV}}, \lambda_{\text{LR}}, \lambda_{\text{albedo}}, \lambda_{\text{SW cloud}}, \lambda_{\text{LW cloud}})_i \tag{1}$$

where λ_i are means of individual feedbacks calculated from either CMIP5 ($\lambda_{i,\text{CMIP5}}$) or CMIP6 ($\lambda_{i,\text{CMIP6}}$). When the RMSD is calculated from the *code weighted* feedback means compared with *simple* means, it is reduced by about 45% from 67 to 37 $\mathrm{mWm^{-2}K^{-1}}$. Therefore, it is possible that a substantial part of the difference in feedbacks between CMIP6 and CMIP5 can be explained by a suitable choice of weighting which takes into account model code dependence. When the RMSD is calculated for *family weighting* (not shown in the plot), the RMSD is approximately the same at 37 $\mathrm{mWm^{-2}K^{-1}}$, but less so for the *model weighting* (reduced to 56 $\mathrm{mWm^{-2}K^{-1}}$), and a slight increase in RMSD is seen for *institute* (increased to 92 $\mathrm{mWm^{-2}K^{-1}}$) and *country* (increased to 82 $\mathrm{mWm^{-2}K^{-1}}$) weighting. This could mean that only the *code*, *family* and to a lesser extent *model weighting* can explain some of the feedback difference between CMIP6 and CMIP5. The result is consistent with the expectation that the *code weighting* is more suitable than the other types of weighting, which are less strongly related to the model code genealogy.

For ECS and ERF, the differences between weighting methods are also substantial – up to about 0.3 K for ECS and 60 $\mathrm{mWm^{-2}}$ for $\mathrm{ERF_{2x}}$ (Fig. 3a, b). In comparison, the difference in *simple* mean between CMIP6 and CMIP5 is 0.47 K in ECS and 114 $\mathrm{mWm^{-2}}$ in $\mathrm{ERF_{2x}}$, and the standard deviation is 0.73 K and 1.06 K in ECS (CMIP5 and CMIP6, resp.) and 390 $\mathrm{mWm^{-2}}$ and 490 $\mathrm{mWm^{-2}}$ in $\mathrm{ERF_{2x}}$ (CMIP5 and CMIP6, resp.). Ensemble mean ECS difference between CMIP6 and CMIP5 becomes almost zero with the *code weighting*, reduced from 0.47 K (*simple* mean) to 0.06 K (*code weighting*), but the difference in $\mathrm{ERF_{2x}}$ is increased from 114 to 186 $\mathrm{mWm^{-2}}$. Thus, it is possible that a weighting method which accounts for model code dependency can explain some of the large difference in ECS between CMIP5 and CMIP6 due to an overrepresentation of models with high ECS in the CMIP6 ensemble.

4.3 Climate feedbacks and sensitivity by model family

We analysed climate feedbacks and sensitivity by model family (Fig. 4). Because model *family weighting* showed results similar to *code weighting* (Sect. 4.2), it should be a good proxy for *code weighting*, while allowing us to separate the values into (potentially clustered) groups. Some model families tend to have similar values of climate feedbacks. This is most apparent in the cloud feedbacks, where differences between models are generally large. The HadAM family of models tend to be closely clustered in all climate feedbacks, despite the comparatively large size of the model family (6 models in the CMIP6 plot). Their total cloud and SW cloud feedback is consistently larger than the mean and their LW cloud feedback is consistently smaller than the mean. The ECMWF family of models (14 models in the CMIP6 plot) have consistently below mean SW cloud feedback, mostly below mean total cloud feedback and almost consistently above mean LW cloud feedback. The CCM family is the largest (17 models in the CMIP6 plot) and also the most varied, showing a large spread between the models in CMIP6, but a small spread in CMIP5. Despite this, they have some characteristic properties, such as in CMIP6 mostly above mean total and SW cloud feedback and below mean LW cloud feedback; in CMIP5 mostly below mean total cloud feedback, but also above mean lapse rate and surface albedo and below mean water vapour feedback. In CMIP6, the UCLA GCM family of models (5 models in the CMIP6 plot) have consistently below mean total and SW cloud feedback and mostly above mean LW cloud feedback.

In terms of ECS, the CCM and ECMWF family of models show a large and quite even spread around the multi-model mean. In CMIP6, the HadAM and IPSL family of models are all more sensitive than the mean, and the UCLA GCM family of models are all less sensitive than the mean.

In summary, some relatively large families of models show consistent properties when it comes to climate feedbacks and ECS, while others show a large spread. This suggests that models in some models families have substantial dependence which translates into clustering of climate feedbacks and ECS. The CCM and ECMWF families are quite diverse, but despite this they show common characteristics in some climate feedbacks.

4.4 Global mean near-surface temperature time series

To analyse the impact of the *code* and model *family weighting* methods on MME statistics, we examine the case of GMST in the *historical*, SSP2-4.5, *abrupt-4xCO2* and *1pctCO2* CMIP6 experiments and the *historical*, RCP4.5, *abrupt-4xCO2* and *1pctCO2* CMIP5 experiments. Figure 5 and 6 show GMST time series in the CMIP6 and CMIP5 experiments (respectively), grouped by model family, as well as *family* and *code weighted* time series. Included are all models which provided the necessary data. While some model families have many members in this analysis, such as CCM (7 to 22 members), ECMWF (3 to 16 members), HadAM (2 to 6 members) and UCLA GCM (1 to 5 members) other families have less than 4 members, and therefore it is harder (or impossible) to assess model spread in the smaller families. The larger families of models such as CCM and ECMWF exhibit a large spread and a middle-of-the-range family mean, although the spread of the ECMWF family in the CMIP5 *historical* + RCP4.5, *abrupt-4xCO2* and *1pctCO2* experiments is relatively narrow. The other larger family HadAM has a relatively small spread in most experiments, consistent with the results of Sect. 4.3. Notably, in the CMIP6 *historical* experiment, HadAM is

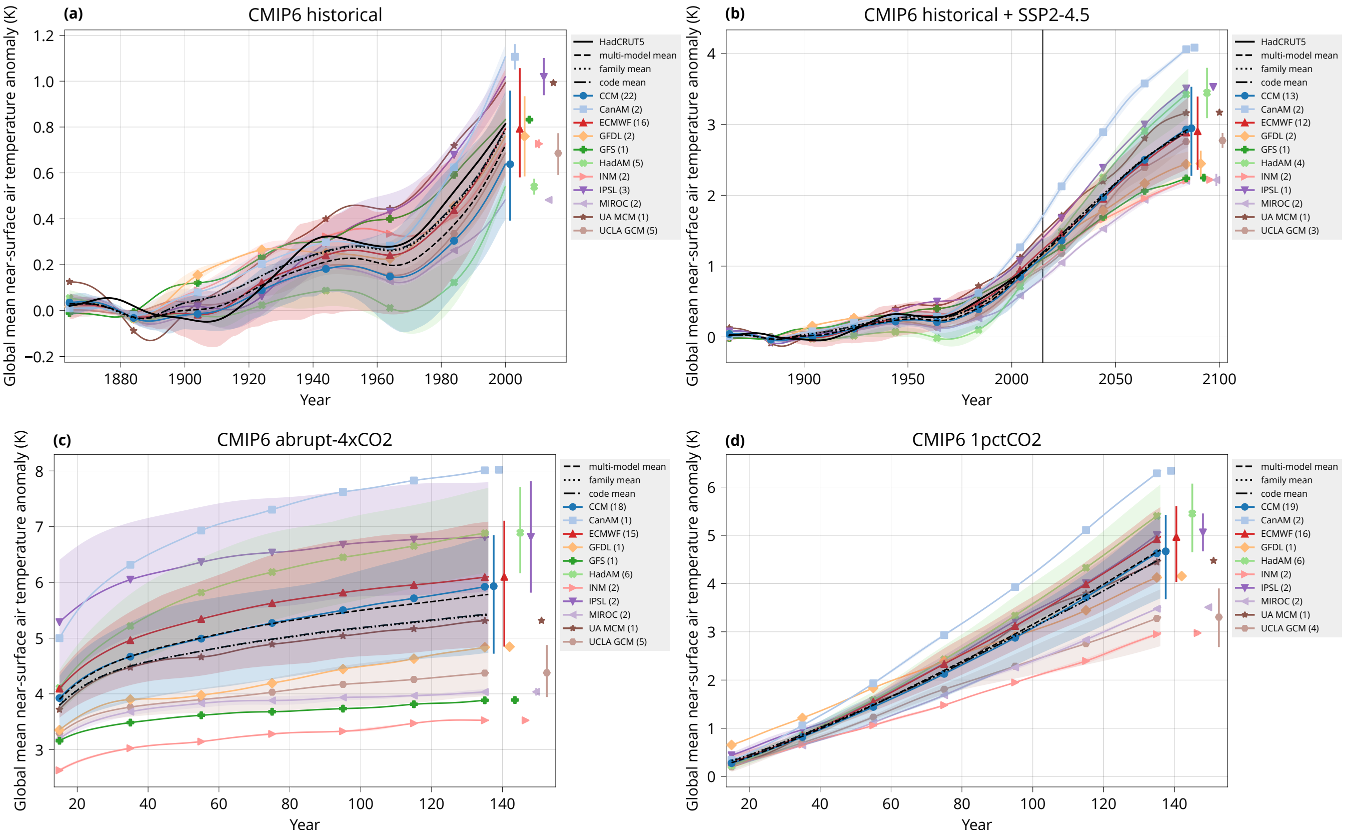

Figure 5. Time series of global mean near-surface temperature in CMIP6 experiments by model family and the *simple* multi-model, *code* and *family* mean (Sect. 3.2). The model family time series are a *simple* mean of models in the family. The time series are smoothed with a Gaussian kernel with a standard deviation of 7 years. The first and the last 14 years of the time series are not shown to avoid artefacts of the smoothing. The values are relative to the mean of the first 30 years of the individual time series in **(a)** and **(b)**, and relative to the mean of the whole individual time series of the *piControl* experiment in **(c)** and **(d)**. Shown are confidence bands representing the 68th percentile range. The vertical divider in the *historical* + SSP2-4.5 plot separates time range of the two experiments. In the legend, the number in the parentheses is the number of models in the family. All CMIP5 and CMIP6 models with necessary data available on the Earth System Grid were included in the plots.

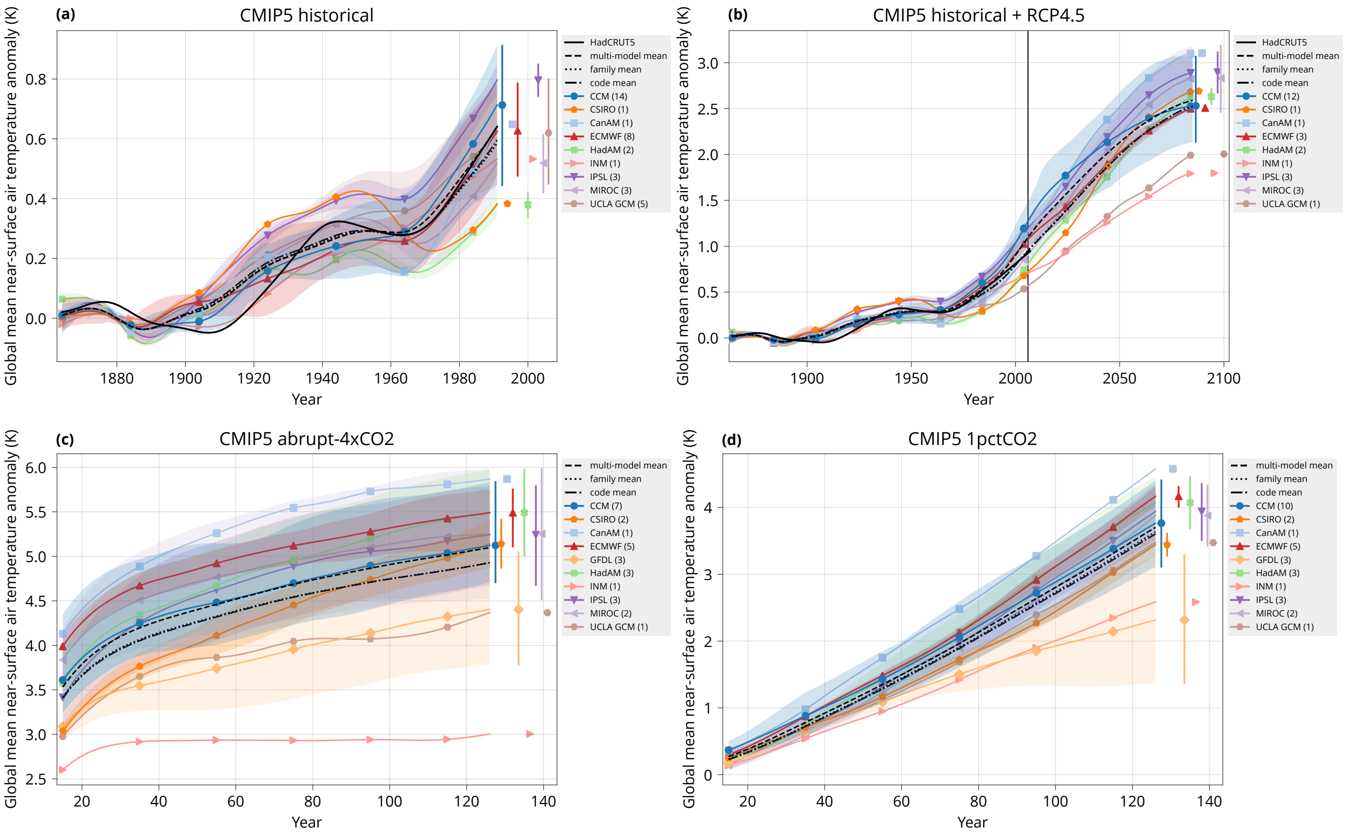

Figure 6. The same as Fig. 5 but for CMIP5, and the RCP4.5 experiment instead of SSP2-4.5.

the coldest of all model families, but becomes the second and third warmest in the rest of the CMIP6 experiments by the end of the simulation. The UCLA GCM family of models have consistently relatively low GMST in the CMIP6 *abrupt-4xCO2* and *1pctCO2* experiments, despite the relatively large size of the group (here 4 to 5 members). Model families like MIROC, INM and CanAM (containing 2 members in the CMIP6 plots, except for CanAM in *abrupt-4xCO2* with only member) have almost no spread in the CMIP6 experiments, suggesting that the two models in each of these model families are very similar. 340

The *family* and *code weighted* GMST time series tend to nearly overlap in all cases, which points to a high degree of outcome similarity between the two types of weighting also noted in the preceding sections. Interestingly, the *family* and *code weighted* mean is warmer than the *simple* multi-model mean in the CMIP6 *historical* experiment (in the CMIP5 *historical* experiment it is slightly colder by the end of the simulation), and it is also more consistent with observations, whereas in the in the *1pctCO2* 345 and *abrupt-4xCO2* experiments it is colder than the *simple* mean (in both CMIP6 and CMIP5). When CMIP6 is compared with CMIP5, model families tended to exhibit similar cold or warm propensity, such as INM, GFDL, UCLA GCM being relatively cold in the non-*historical* experiments, and CanAM, HadAM, IPSL relatively warm. This suggests that model families tend to maintain their climate sensitivity inclination across model generations.

5 Discussion and conclusions

We mapped the code genealogy of 167 models in and related to CMIP3, CMIP5 and CMIP6 with a focus on the atmospheric component and the atmospheric physics. We showed that all models can be grouped into 14 model families based on code inheritance, although in some models large amounts of code may have been replaced, and therefore they are only weakly related to other models in the same family. In addition, we mapped the institute and country of origin of the models. Some model families such as CCM, ECMWF and HadAM were particularly large. The CCM-derived models were forked very commonly internationally, most likely due to the open availability of the code. The IFS/ARPEGE (licensed) code was the basis for many European models. The HadGEM code was shared internationally within a consortium. Together, these three large model families dominated CMIP6, accounting for 70% of all model runs, and increase from about 50% represented by the three largest model families in CMIP3 and CMIP5. Based on the code genealogy, we developed a *code weighting* method, whose aim was to more fairly weigh code-related models than a *simple* multi-model mean, thus mitigating structural model dependence in MMEs. We showed that when applied on CMIP5 and CMIP6, the *code* and *family weighting* produced substantial differences in the climate feedbacks, sensitivity and forcing, especially the cloud feedbacks (total, shortwave and longwave), ECS and ERF_{2x} – relative to the difference in *simple* mean between CMIP6 and CMIP5, and the standard deviation of the quantities in CMIP5 and CMIP6. The *code* and *family weighting* methods showed very similar results. The *code* and *family weighting* seem to be able to reconcile some of the difference between CMIP6 and CMIP5 – about 45% RMSD reduction in climate feedbacks, and 87% RMSD reduction in ECS under the *code weighting*. This suggests that increased contributions from many code-related models in CMIP6 compared to CMIP5 were able to substantially affect the *simple* multi-model mean. Applying these methods to analyse climate feedbacks, sensitivity and forcing by model family revealed that models in some families gave closely similar results (HadAM and UCLA GCM), and others sometimes had relatively wide spread but consistently above-mean or below-mean values (ECMWF and CSM). This suggests that code similarity at least in same cases translates to similarities in climate properties. Lastly, we analysed GMST time series in four CMIP6 and CMIP5 experiments, and showed that models in some larger families (HadAM, and in some cases ECMWF) have similar GMST. The *family* and *code weighting* showed very similar results – more warming than the *simple* mean (and closer to observations) in the CMIP6 *historical* experiment and less warming in the CMIP6 *1pctCO2* and *abrupt-4xCO2* experiments. This suggests that they can be successful in balancing the effect of the over-representation of model families with many similar models like HadAM. Model families tend to exhibit tendencies of greater or smaller warming than the MME mean in response to CO_2 increase across the CMIP generations.

Our analysis had some limitations. We did not make an attempt to quantify model code independence from their parent models, because there is not enough publicly available information on the source code. Even if the source code were available, an objective quantification of code independence would require a sophisticated new method of code analysis. Some models have code bases which are more independent from their parent models than others. As a result, some model families might have members which are almost code-independent from the rest of the family.

We do not generally argue against the use of *simple* multi-model means, or model output and performance weighting methods, but see the presented weighting methods as complementary to the established methods. *Simple* means will likely continue to represent a useful default option (as used, for example, in parts of AR6), but other weighting methods may be increasingly important due to model duplication in MMEs. It is possible that weighting methods based on model structure can capture these duplicities better than methods based on model output. We suggest the family weighting, or a similar technique based on selecting a number of ‘independent’ model branches from the model code genealogy, as a useful and easy to implement method of weighting for MME studies, especially if there is an expectation that model duplication is affecting the results.

The presented model code genealogy (Fig. 2) can be further extended as more models become available in the next CMIP phases. We provide the source of this figure in the supplement, and all related code and data are in the supplementary code under an open source license.

Our results can facilitate MME assessments, which depend on the knowledge of model code relations. They provide a complementary approach to the model output dependence methods presented in previous studies. We showed that as expected, code-related models tend to have related climate characteristics and GMST, and this may help to explain some of the difference between CMIP5 and CMIP6. Certain model families stand out in terms of ECS or climate feedbacks, which can help in understanding model differences. This is especially important given that the model spread in ECS and some climate feedbacks increased in CMIP6 relative to CMIP5. A useful method of accounting for dependencies among models is weighting model families equally, which has the benefit of being simpler to achieve than code weighting. This can be readily deployed in MME assessments if a more fair model weighting is desired.

Appendix A: Model code weight calculation

Statistical weights in model *code weighting* are calculated using the model code genealogy in Fig. 2. The weights are calculated for a set of models of interest, i.e. those models or their runs (configuration or resolution) which are present in an MME. Definitions:

1. *Node* is a single model (AGCM, AOGCM or ESM). It can comprise multiple model runs (configurations or resolutions) submitted to CMIP. Nodes can have one or more parent and child nodes.

2. *Model run* is a specific model configuration or resolution submitted to CMIP. Some models only have one run in CMIP.

3. *Group* is a set of nodes with the same model name but different version numbers. In Fig. 2, these are connected with horizontal arrows. Group ancestors are all node ancestors of all nodes in the group.

4. *Root nodes* are nodes which do not have have any ancestors. These are the top-level nodes marked with a thick outline in Fig. 2.

5. *Root groups* are groups which contain a root node.

6. *Active nodes* and *active model runs* are those which are included in the set of models of interest, i.e. models for which weights are to be calculated.

7. *Active groups* are groups which contain at least one active node.
8. *Child node* and *child group* is a direct descendant of its *parent node* or *parent group*.
9. *Descendant* of a node or group is a direct or indirect (more than one level deep) descendant of the node or group.

Algorithm steps (note that the definition of x and n varies by step):

1. Groups and nodes which are not active and have no active descendants are removed from the tree.
2. All nodes and groups are assigned a weight of zero.
3. All root groups are given the same weight equal to $1/n$, where n is the number of root groups.
4. For all groups which have already inherited weight from all of their ancestors (or have no ancestors) and are not marked as done, their child groups inherit weight. If the parent group is active, each child group's weight is incremented by $1/(n+1)$, where n is the number of child groups, and the parent group's weight is set to $1/(n+1)$. If the parent group is not active, each child group's weight is incremented by $1/n$, and the parent group's weight is set to zero. The parent group is marked as done.
5. If all groups are marked as done, continue with Step 6. Otherwise, go back to Step 4.
6. Within each group, active nodes are given weight equal to x/n, where x is the weight of the group and n is the number of active nodes in the group.
7. For each node, active model runs of the node are given weight equal to x/n, where x is the weight of the node and n is the number of active model runs.

Code and data availability. Our processing and visualisation code is available on GitHub (Kuma, 2022b) and Zenodo (Kuma, 2022a). The CMIP5 and CMIP6 model output data are publicly available on the Earth System Grid Federation websites (CMIP5, 2022; CMIP6, 2022).

Author contributions. PK – conceptualisation, formal analysis, investigation, methodology, software, visualisation and writing (original draft preparation). FB – conceptualisation, funding acquisition, methodology, project administration, supervision and writing (review & editing). AJ – conceptualisation, methodology, software and writing (review & editing).

Competing interests. The authors declare no competing interests.

Acknowledgements. We would like to acknowledge funding from the FORCeS project: ‘Constrained aerosol forcing for improved climate projections’ (https://forces-project.eu), funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement number 821205. We acknowledge the World Climate Research Programme (WCRP), the Climate Model Intercomparison Project (CMIP), the Earth System Grid Federation (ESGF) and the climate modelling groups for providing the model output data. We acknowledge the use of open source software Python, NumPy, Matplotlib, SciPy, Inkscape and Devuan GNU+Linux.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth System Dynamics, 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Alexander, K. and Easterbrook, S. M.: The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations, Geoscientific Model Development, 8, 1221–1232, https://doi.org/10.5194/gmd-8-1221-2015, 2015.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Rashid, H., Uotila, P., Hirst, A., Kowalczyk, E., Golebiewski, M., Sullivan, A., Yan, H., Hannah, N., Franklin, C., Sun, Z., Vohralik, P., Watterson, I., Zhou, X., Fiedler, R., Collier, M., Ma, Y., Noonan, J., Stevens, L., Uhe, P., Zhu, H., Griffies, S., Hill, R., Harris, C., and Puri, K.: The ACCESS coupled model: description, control climate and evaluation, Australian Meteorological and Oceanographic Journal, 63, 41–64, https://doi.org/10.1071/ES13004, 2013.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Climate dynamics, 41, 885–900, https://doi.org/10.1007/s00382-012-1610-y, 2013.

Boé, J.: Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity, Geophysical Research Letters, 45, 2771–2779, https://doi.org/10.1002/2017GL076829, 2018.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., and Sanderson, B. M.: Statistical significance of climate sensitivity predictors obtained by data mining, Geophysical Research Letters, 41, 1803–1808, https://doi.org/10.1002/2014GL059205, 2014.

CMIP3: WCRP Coupled Model Intercomparison Project phase 3 (CMIP3), https://esgf-node.llnl.gov/projects/cmip3/, last access: 1 August 2022, 2022.

CMIP5: WCRP Coupled Model Intercomparison Project 5 (CMIP5), https://esgf-node.llnl.gov/projects/cmip5/, last access: 1 August 2022, 2022.

CMIP6: WCRP Coupled Model Intercomparison Project (Phase 6), https://esgf-node.llnl.gov/projects/cmip6/, last access: 1 August 2022, 2022.

Edwards, P. N.: Chapter 2 - A Brief History of Atmospheric General Circulation Modeling, in: General Circulation Model Development, edited by Randall, D. A., vol. 70 of *International Geophysics*, pp. 67–90, Academic Press, https://doi.org/10.1016/S0074-6142(00)80050-9, 2000a.

Edwards, P. N.: Atmospheric General Circulation Modeling: A Participatory History, http://pne.people.si.umich.edu/sloan/mainpage.html, last access: 12 August 2022, 2000b.

Edwards, P. N.: History of climate modeling, WIREs Climate Change, 2, 128–139, https://doi.org/10.1002/wcc.95, 2011.

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

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nature Climate Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., and Zhang, H.: 2021: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, chap. The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, pp. 923–1054, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.009, 2021.

Gjermundsen, A., Nummelin, A., Olivié, D., Bentsen, M., Seland, Ø., and Schulz, M.: Shutdown of Southern Ocean convection controls long-term greenhouse gas-induced warming, Nature Geoscience, 14, 724–731, https://doi.org/10.1038/s41561-021-00825-x, 2021.

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., Abeshu, G., Anantharaj, V., Asay-Davis, X. S., Bader, D. C., Baldwin, S. A., Bisht, G., Bogenschutz, P. A., Branstetter, M., Brunke, M. A., Brus, S. R., Burrows, S. M., Cameron-Smith, P. J., Donahue, A. S., Deakin, M., Easter, R. C., Evans, K. J., Feng, Y., Flanner, M., Foucar, J. G., Fyke, J. G., Griffin, B. M., Hannay, C., Harrop, B. E., Hoffman, M. J., Hunke, E. C., Jacob, R. L., Jacobsen, D. W., Jeffery, N., Jones, P. W., Keen, N. D., Klein, S. A., Larson, V. E., Leung, L. R., Li, H.-Y., Lin, W., Lipscomb, W. H., Ma, P.-L., Mahajan, S., Maltrud, M. E., Mametjanov, A., McClean, J. L., McCoy, R. B., Neale, R. B., Price, S. F., Qian, Y., Rasch, P. J., Reeves Eyre, J. E. J., Riley, W. J., Ringler, T. D., Roberts, A. F., Roesler, E. L., Salinger, A. G., Shaheen, Z., Shi, X., Singh, B., Tang, J., Taylor, M. A., Thornton, P. E., Turner, A. K., Veneziani, M., Wan, H., Wang, H., Wang, S., Williams, D. N., Wolfram, P. J., Worley, P. H., Xie, S., Yang, Y., Yoon, J.-H., Zelinka, M. D., Zender, C. S., Zeng, X., Zhang, C., Zhang, K., Zhang, Y., Zheng, X., Zhou, T., and Zhu, Q.: The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution, Journal of Advances in Modeling Earth Systems, 11, 2089–2129, https://doi.org/10.1029/2018MS001603, 2019.

Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S., Lautenschlager, M., Morgan, M., Murphy, S., and Taylor, K. E.: Documenting Climate Models and Their Simulations, Bulletin of the American Meteorological Society, 94, 623–627, https://doi.org/10.1175/BAMS-D-11-00035.1, 2013.

Haughton, N., Abramowitz, G., Pitman, A., and Phipps, S. J.: Weighting climate model ensembles for mean and variance estimates, Climate dynamics, 45, 3169–3181, https://doi.org/10.1007/s00382-015-2531-3, 2015.

Jun, M., Knutti, R., and Nychka, D. W.: Spatial Analysis to Quantify Numerical Model Bias and Dependence, Journal of the American Statistical Association, 103, 934–947, https://doi.org/10.1198/016214507000001265, 2008a.

Jun, M., Knutti, R., and Nychka, D. W.: Local eigenvalue analysis of CMIP3 climate model errors, Tellus A: Dynamic Meteorology and Oceanography, 60, 992–1000, https://doi.org/10.1111/j.1600-0870.2008.00356.x, 2008b.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, Journal of Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophysical Research Letters, 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.

Krishnan, R., Swapna, P., Choudhury, A. D., Narayansetti, S., Prajeesh, A. G., Singh, M., Modi, A., Mathew, R., Vellore, R., Jyoti, J., Sabin, T. P., Sanjay, J., and Ingle, S.: The IITM Earth System Model (IITM ESM), https://doi.org/10.48550/ARXIV.2101.03410, 2021.

Kuma, P.: Code accompanying the manuscript "Climate model code genealogy and its relation to climate feedbacks and sensitivity", https://doi.org/10.5281/zenodo.7407118, 2022a.

Kuma, P.: Code accompanying the manuscript "Climate model code genealogy and its relation to climate feedbacks and sensitivity", https://github.com/peterkuma/model-code-genealogy-2022/, last access: 6 December 2022, 2022b.

Lenhard, J. and Winsberg, E.: Holism, Entrenchment, and the Future of Climate Model Pluralism, Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 41, 253–262, https://doi.org/10.1016/j.shpsb.2010.07.001, 2010.

Lynch, P.: The origins of computer weather prediction and climate modeling, Journal of Computational Physics, 227, 3431–3444, https://doi.org/10.1016/j.jcp.2007.02.034, predicting weather, climate and extreme events, 2008.

Masson, D. and Knutti, R.: Climate model genealogy, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL046864, 2011.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., eds.: 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, Cambridge, United Kingdom, in press, 2021.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., and Taylor, K. E.: THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research, Bulletin of the American Meteorological Society, 88, 1383–1394, https://doi.org/10.1175/BAMS-88-9-1383, 2007.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models, Science Advances, 6, eaba1981, https://doi.org/10.1126/sciadv.aba1981, 2020.

Mendlik, T. and Gobiet, A.: Selecting climate simulations for impact studies based on multivariate patterns of climate change, Climatic change, 135, 381–393, https://doi.org/10.1007/s10584-015-1582-0, 2016.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of state calculations by fast computing machines, The journal of chemical physics, 21, 1087–1092, 1953.

Pennell, C. and Reichler, T.: On the Effective Number of Climate Models, Journal of Climate, 24, 2358–2367, https://doi.org/10.1175/2010JCLI3814.1, 2011.

Pulkkinen, K., Undorf, S., Bender, F., Wikman-Svahn, P., Doblas-Reyes, F., Flynn, C., Hegerl, G. C., Jönsson, A., Leung, G.-K., Roussos, J., Shepherd, T. G., and Thompson, E.: The value of values in climate science, Nature Climate Change, 12, 4–6, https://doi.org/10.1038/s41558-021-01238-9, 2022a.

Pulkkinen, K., Undorf, S., and Bender, F. A.-M.: Values in climate modelling: testing the practical applicability of the Moral Imagination ideal, European Journal for Philosophy of Science, 12, 68, https://doi.org/10.1007/s13194-022-00488-4, 2022b.

Remmers, J. O., Teuling, A. J., and Melsen, L. A.: Can model structure families be inferred from model output?, Environmental Modelling & Software, 133, 104 817, https://doi.org/10.1016/j.envsoft.2020.104817, 2020.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C.: Probabilistic programming in Python using PyMC3, PeerJ Computer Science, 2, e55, https://doi.org/10.7717/peerj-cs.55, 2016.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties, Journal of Climate, 28, 5150–5170, https://doi.org/10.1175/JCLI-D-14-00361.1, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, Journal of Climate, 28, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1, 2015b.

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B., Fisher, R. A., and Knutti, R.: The potential for structural errors in emergent constraints, Earth System Dynamics, 12, 899–918, https://doi.org/10.5194/esd-12-899-2021, 2021.

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., and Eyring, V.: Emergent constraints on equilibrium climate sensitivity in CMIP5: do they hold for CMIP6?, Earth System Dynamics, 11, 1233–1258, https://doi.org/10.5194/esd-11-1233-2020, 2020.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Reviews of Geophysics, 58, e2019RG000 678, https://doi.org/10.1029/2019RG000678, e2019RG000678 2019RG000678, 2020.

Steinschneider, S., McCrary, R., Mearns, L. O., and Brown, C.: The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning, Geophysical Research Letters, 42, 5014–5044, https://doi.org/10.1002/2015GL064529, 2015.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Touzé-Peiffer, L., Barberousse, A., and Le Treut, H.: The Coupled Model Intercomparison Project: History, uses, and structural effects on climate research, WIREs Climate Change, 11, e648, https://doi.org/10.1002/wcc.648, 2020.

Voosen, P.: 'Hot' climate models exaggerate Earth impacts, Science (New York, NY), 376, 685–685, https://doi.org/10.1126/science.adc9453, 2022.

Wang, C., Soden, B. J., Yang, W., and Vecchi, G. A.: Compensation Between Cloud Feedback and Aerosol-Cloud Interaction in CMIP6 Models, Geophysical Research Letters, 48, e2020GL091 024, https://doi.org/10.1029/2020GL091024, e2020GL091024 2020GL091024, 2021.

Williams, J., Morgenstern, O., Varma, V., Behrens, E., Hayek, W., Oliver, H., Dean, S., Mullan, B., and Frame, D.: Development of the New Zealand Earth system model, Weather and Climate, 36, 25–44, https://doi.org/10.2307/26779386, 2016.

Winsberg, E.: Values and Uncertainties in the Predictions of Global Climate Models, Kennedy Institute of Ethics Journal, 22, 111–137, https://doi.org/10.1353/ken.2012.0008, 2012.

Zelinka, M. D.: GitHub repository mzelinka/cmip56_forcing_feedback_ecs, https://github.com/mzelinka/cmip56_forcing_feedback_ecs, last access: 3 August 2022, 2022.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of Higher Climate Sensitivity in CMIP6 Models, Geophysical Research Letters, 47, e2019GL085 782, https://doi.org/10.1029/2019GL085782, e2019GL085782 10.1029/2019GL085782, 2020.