# 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) and Bolin Centre for Climate Research, Stockholm University, Stockholm, SE-106 91, Sweden

**Key Points:**

- We reconstruct a code genealogy of 167 climate models with a focus on the atmospheric component and atmospheric physics.
- All models originate from 12 main model families, and models in the same family often have similar climate feedbacks and sensitivity.
- Proposed ancestry and family weighting can partly reconcile differences in means between the Coupled Model Intercomparison Project phases.

Corresponding author: Peter Kuma, peter.kuma@misu.su.se

**Abstract**

Contemporary general circulation models (GCMs) and Earth system models (ESMs) are developed by a large number of modeling groups globally. They use a wide range of representations of physical processes, allowing for structural (code) uncertainty to be partially quantified with multi-model ensembles (MMEs). Many models in the MMEs of the Coupled Model Intercomparison Project (CMIP) have a common development history due to sharing of code and schemes. This makes their projections statistically dependent and introduces biases in MME statistics. Previous research has focused on model output and code dependence, and model code genealogy of CMIP models has not been fully analyzed. We present a full reconstruction of CMIP3, CMIP5 and CMIP6 code genealogy of 167 atmospheric models, GCMs, and ESMs (of which 114 participated in CMIP) based on the available literature, with a focus on the atmospheric component and atmospheric physics. We identify 12 main model families. We propose family and ancestry weighting methods designed to reduce the effect of model structural dependence in MMEs. We analyze weighted effective climate sensitivity (ECS), climate feedbacks, forcing, and global mean near-surface air temperature, and how they differ by model family. Models in the same family often have similar climate properties. We show that weighting can partially reconcile differences in ECS and cloud feedbacks between CMIP5 and CMIP6. The results can help in understanding structural dependence between CMIP models, and the proposed ancestry and family weighting methods can be used in MME assessments to ameliorate model structural sampling biases.

## Plain Language Summary

Contemporary global climate models are developed by a large number of modeling groups internationally. Commonly, projections from multiple models are used together to calculate multi-model means and quantify uncertainty. Because many of the models share parts of their computer code, algorithms and parametrization schemes, they are not independent. Overrepresented models can cause biases in multi-model means, and uncertainty may be underestimated if model dependence is not taken into account. We document a full code genealogy of 167 models, of which 114 participated in the Coupled Model Intercomparison Project (CMIP) phases 3, 5, and 6, with a focus on the atmospheric component. We identify 12 main model families. We show that models in the same family often have similar estimates of key climate properties. We propose statistical weighting methods based on the model family and code relationship, and show that they can reconcile some of the difference in results between the two most recent CMIP phases. The weighting methods or a selection of independent models based on the genealogy can be used in model assessment studies to reduce the effects of model dependence.

## 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 development of numerical weather prediction (NWP) models (Lynch, 2008). Intercomparison between climate models dates back to the late 1980s when the Atmospheric Model Intercomparison Project (AMIP) started comparing atmospheric models under standardized conditions and model output (Touzé-Peiffer et al., 2020). This was followed by the Coupled Model Intercomparison Project (CMIP) phase 1 and 2 in 1996 and 1997, respectively, which 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 data set 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 & Easterbrook, 2015). These components may be divided into subcomponents, modules or schemes representing various physical parametrizations, such as radiative transfer in the atmospheric component. Components and subcomponents can sometimes 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 modeling group as well as between groups internationally (in the following text we will use the terms "modeling group" and "institute", the latter being common in the context of CMIP, interchangeably). Alexander and Easterbrook (2015) directly analyzed the source code of model components, showing significant sharing of components between models thanks to their highly modular nature. Furthermore, parametrizations documented in literature were implemented in a variety of models, meaning that they use many of the same parametrizations 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 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 common in software development. This is also the case with climate models, where different modeling groups base their work on forking of an existing model from the same or a different modeling 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 the proliferation of code which produces the best results, which could otherwise arise if all model code were openly available. As discussed below, what constitutes "the best results" may be difficult to quantify and is not guaranteed to coincide with the best projections. Guilyardi et al. (2013) initiated better model and experiment metadata collection within CMIP5 in order 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 categorized 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 data sets for validation, and implementing similar parametrizations. Some authors have therefore called this MME an "ensemble of opportunity" (Masson & Knutti, 2011; Knutti et al., 2013; Sanderson et al., 2015a; Boé, 2018), since the inclusion is based on the intent of a modeling group to participate rather than objective selection criteria. If model dependence is not taken into account, the calculation of means, variance, and uncertainty can be biased, and spurious correlations (such as in emergent constraints) can arise in an MME (Caldwell et al., 2014; Sanderson et al., 2021). Remmers et al. (2020) investigated whether model code genealogy can be inferred from model output [also investigated earlier by Knutti et al. (2013) and discussed below]. Using a modular modeling framework, they generated a model ensemble of hydrological models by sampling the model "hypothesis space" [as defined in Remmers et al. (2020)] 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 the 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. Increasing the number of ensemble models did not substantially increase the effective number of models. Sanderson et al. (2015b) reached a similar conclusion, and found that the number of independent models calculated based on the model output in CMIP5 is much smaller than the total.

The simplest approach to analyzing 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, 2008b; Masson & Knutti, 2011; Knutti et al., 2013; Bishop & Abramowitz, 2013; Sanderson et al., 2015a; Haughton et al., 2015; Mendlik & Gobiet, 2016), while a focus on code dependence has been relatively rare (Alexander & 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 a hierarchical clustering of model output. They found that models from the same institute were much closer in their model output than other models, and contemplated that output similarity could be used for model weighting or selection to eliminate biases due to near duplicate models. A more simple approach is "institutional democracy", where one model per modeling group is selected, and "component democracy", where models are selected to represent different model components (Abramowitz et al., 2019). Edwards (2000a, 2000b, 2000c, 2011, 2013) described the early to modern history of climate modeling and constructed a partial "family tree" of atmospheric GCMs based on their code heritage. Another account on early climate modeling was given by Arakawa (2000). Boé (2018) summarized institute, 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 analyzed. 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 relative to observations (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 the 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 permitting only one model per institute in an MME 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 developed alongside HadGEM/UKESM) was developed to reduce Southern Ocean biases (Williams et al., 2016); the Indian Institute of Tropical Meteorology ESM (IITM ESM) has 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) aims to support the U.S. energy sector decisions (Golaz et al., 2019). Weighting models by errors relative to observations (performance weighting) 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 projected change (or trend) of the variables under a climate scenario (Jun et al., 2008a; Zelinka, 2022; Kuma et al., 2022). Thus, a model's performance in future climate projections cannot be fully inferred from its performance in present-day and historical climate. Performance weighting can also favor 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 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 are tuned less to represent present-day or historical observations. Knutti (2010) provides a high-level discussion of the topic of model democracy, uncertainty, weighting, evaluation, calibration and tuning in the context of decision making.

Apart from explicit model weighting or selection choices, seldomly recognized implicit choices based on values (other than widely acknowledged epistemic values such as openness, objectivity, evidence, and impartiality) influence model development, evaluation, selection, weighting, interpretation, and communication of results (Pulkkinen, Undorf, Bender, Wikman-Svahn, et al., 2022; Pulkkinen, Undorf, & Bender, 2022; Lenhard & Winsberg, 2010; Winsberg, 2012; Undorf et al., 2022). The climate system is too complex to be captured by models perfectly. Some of the limitations stem from limited computational resources, uncertainty about how to represent processes at a coarse level through parametrizations, and a lack of observational data. Thus, model construction necessitates and is affected by decisions regarding a variety of compromises. Traditionally, a pursuit of purely knowledge-oriented science has been desired in order to avoid conclusions distorted by scientists' views, values and interests. However, some authors emphasize that purely knowledge-oriented construction of climate models is impossible because of decisions involved in the model development (Parker & Winsberg, 2018; Parker, 2020; Jebeile & Crucifix, 2021; Morrison, 2021). These decisions can be driven by not only the desire for creating an unbiased objective representation of the climate system, but also by purposes, views, values, interests and limitations. They include for example a specific focus on modeling a certain geographical region and quantities of interest, the availability of validation data influenced by locations of observations, compromises regarding what errors are permissible, types of tuning (Schmidt et al., 2017), decisions involved in earlier versions of the same model or ancestral models resulting in inherited values, limited knowledge and time of the researchers, and limited resources. In turn, they can also perpetuate certain types of societal biases against traditionally understudied and underrepresented regions. Rarely are such decisions or values and interests which drive them explicitly acknowledged, which makes it difficult to quantify their impact on MMEs. Although less acknowledged, interests can also include reasons for pursuing certain research or development which are not driven by practical reasons but by curiosity. In a broader view, the development of climate models has aspects of iterative development, inheritance, recombination, cooperation, competition and filling of different niches. In this way, it can be considered a collective optimization process with the goal of describing the important and diverse properties of the climate system (as considered by various actors) through pluralism in the face of limited knowledge and computational resources, both of which also keep changing.

We can define the structure (code) of a model as based on a set of hypotheses about reality as well as computational realizations 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, and it is not 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. Here, we define a "true representation" in limited terms as a pragmatically-oriented conceptualization of the Earth system, which for example might not include the anthroposphere as commonly externalized in CMIP models through scenarios. Models can be similar in their 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. We note that some authors advocate against a value-free ideal to which models should converge (Parker & Winsberg, 2018; Parker, 2020).

As a conceptual model (Figure 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 conceptualized theoretically by a probability density function (PDF). Ideally, models in an MME are independent samples from this PDF (Figure 1a). In actual MMEs (Figure 1b), however, models are dependent and tend to be clustered together for reasons incompatible with the PDF, such as the 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 organizational 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, the value with the maximum likelihood or the mean are unlikely to coincide with the true physical representation, because such a PDF only represents our belief that a given physical representation is true, which is 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. Other aspects not considered in Figure 1 are that our knowledge about the climate system is shaped by various decisions such as which parts of the climate system have been considered interesting to study or observe, and individual models are also affected by such decisions during their development. As mentioned above, some models even have a particular explicitly stated purpose, such as ACCESS-CM, E3SM, IITM ESM and NZESM. The consequence of this is that models are not only biased samples of the PDF due to code dependence, but also due to value and interest-based decisions. For the same reasons they can also converge or diverge.

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 other models 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

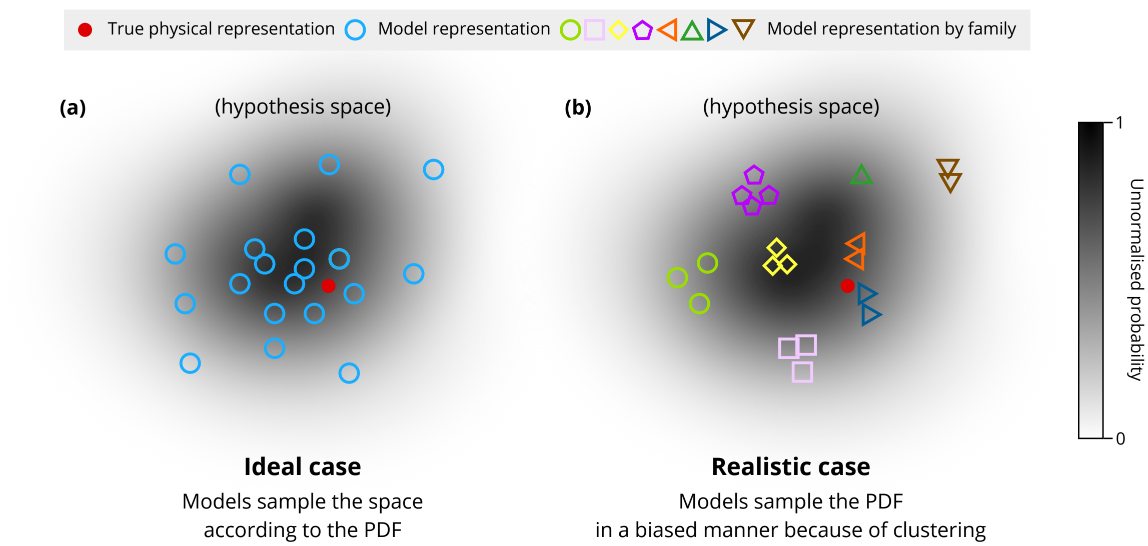


**Figure 1.** A theoretical illustrative example of model sampling of the model hypothesis space (model structural uncertainty), representing realizations of physical climate processes (model structure). The shading indicates a probability density function (PDF) quantifying our collective belief that 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. They might also deviate from the PDF for a number of other reasons. 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, the hypothesis space is visualized in a 2-dimensional space. In reality, this space has a large number of dimensions and the PDF might not be symmetric. Model marker colors (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.

of climate physics (as defined in the limited terms above). Dependence 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 (Masson-Delmotte et al., 2021). 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 it comes to code sharing between modeling groups. 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 insights into 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 available in literature as well as CMIP model metadata and online resources of modeling groups, but has not been systematically organized across CMIP phases. When determining code relations, our focus is on the atmospheric component and atmospheric physics due to the fact that they are currently the main source of model uncertainty in climate sensitivity, dominated by cloud feedback (Wang et al., 2021a; Forster et al., 2021; Zelinka et al., 2020). Steinschneider et al. (2015) also identified the atmospheric component as being a particularly important factor determining the similarity of climate projections of temperature and precipitation between models. However, other model components such as the ocean 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 whether 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 analyze whether model families differ significantly in their predictions from other model families and a simple multi-model mean.

## 3 Data and 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). The equivalent data from CMIP3 were not analyzed here, but we include all CMIP3 models in the model code genealogy. We used historical global temperature data from the Hadley Centre/Climatic Research Unit global surface temperature dataset version 5 (HadCRUT5) (Morice et al., 2021) obtained from the Met Office Hadley Centre (2022). In order to analyze model code genealogy, we performed a broad literature survey, complemented by CMIP model metadata and information available online, particularly modeling groups' websites. 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 thus necessary for reconstructing the genealogy. The model genealogy information, including related references, is also available in Table S1. Along with relations between models, we identified the model institute, the country where the institute resides, and the model family (defined by the oldest ancestral model in the genealogy). Model parameters such as ECS, TCR, 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:

1. *Simple weighting.* Every model run is given equal weight. By "model run" we mean a model resolution or configuration (as listed in Table S1 in the columns *CMIP3/5/6 names*), not multiple simulations performed with the same model but different initial conditions.
2. *Family weighting.* Model families, defined as a complete branch as shown in Figure 2 (discussed later in section 4.1), were given equal weight. This weight was further subdivided equally between models within the family.
3. *Institute weighting.* Model institutes, as shown in Figure 2 as labels on grey areas, were given equal weight. This weight was further subdivided equally between models within the institute.
4. *Country weighting.* Model host countries, as shown in Figure 2 as labels on grey areas, were given equal weight. This weight was further subdivided equally between models of the same country.
5. *Ancestry weighting.* The oldest ancestor models (marked with a thick outline in Figure 2) were given equal weight. This weight was subdivided gradually through branches to descendant models. This method is described in detail in Appendix Appendix A.
6. *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 *italicized*.

### 3.3 Statistical Significance

Statistical significance in climate feedbacks, sensitivity, and forcing in section 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

Figure 2 presents a graph of model code genealogy based on available literature including all CMIP3, CMIP5 and CMIP6 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 together 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 Figure 2 and also listed in Table S2). The models come from 38 different institutes or institute groups and 15 different countries. Institutes are based on the *institute* attribute of the CMIP data sets (CMIP3, 2022; CMIP5, 2022; CMIP6, 2022) for CMIP models and reference publications or online resources for other models, separated by a slash if multiple institutes 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 recognize 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. Model counts per model family, country, and institute in each CMIP phase are listed in Table S2.

We make 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 early stages of the development in the 1980s (Edwards, 2011), and the separate CCM and ECMWF model families are much larger and more diverse than the other model families. 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 participating in CMIP, while others are much larger, e.g. CCM with 28 models and ECMWF with 23 models in CMIP (here by "model" we mean the main model as in Figure 2 rather than model runs in CMIP). In terms of model runs, CCM, ECMWF, and HadAM are particularly numerously represented in CMIP6 with 32, 27, and 12 model runs, amounting to about 70% of the entire CMIP6 MME (Table S2). This means that there is a strongly 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 modeling groups 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 modeling groups internationally. Other models with private code are used much more narrowly, such as CanAM, CSIRO, IPSL or INM, which are only used by their own modeling group (and possibly a few col-

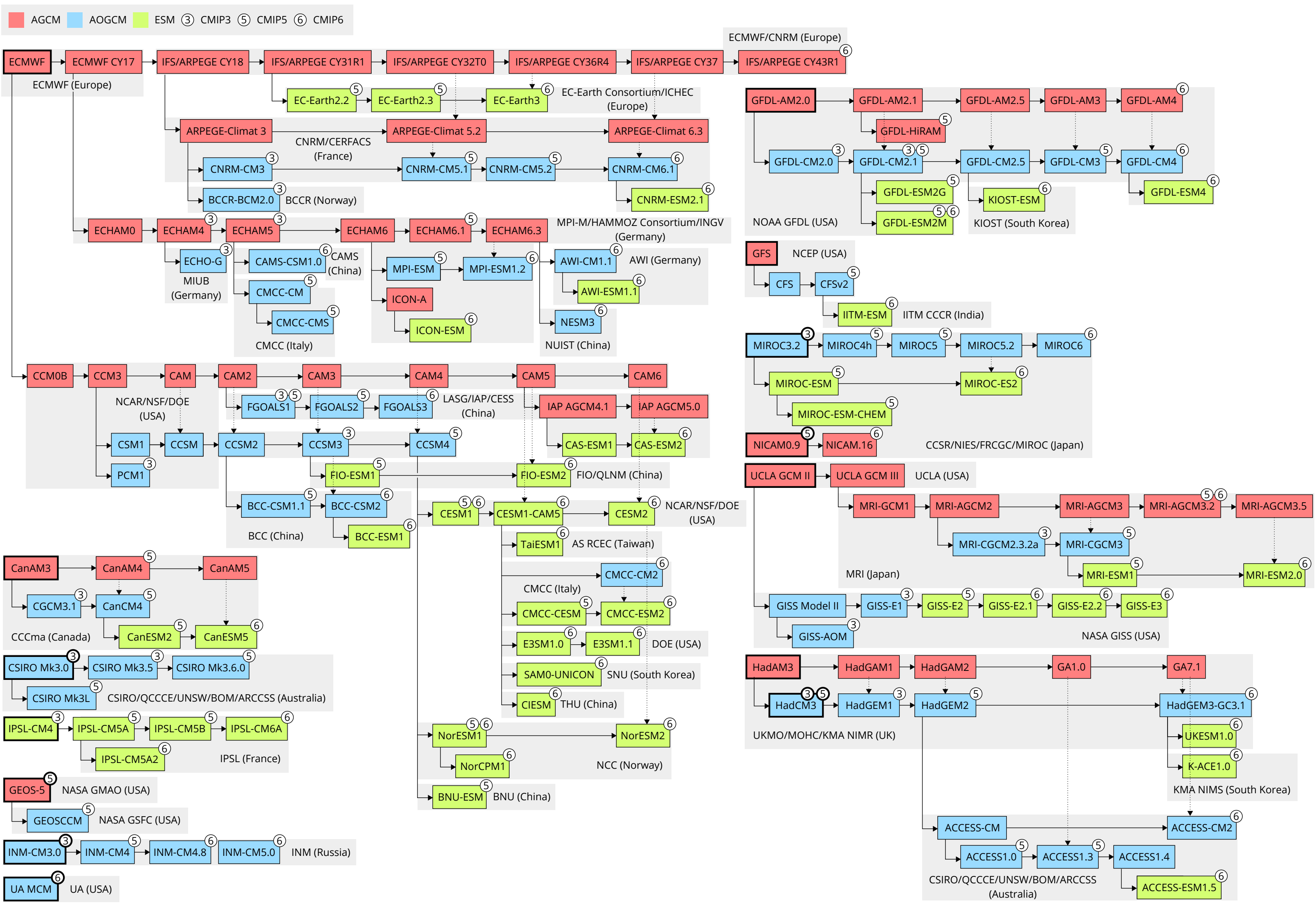


**Figure 2.** Model code genealogy of models participating in the Coupled Model Intercomparison Project (CMIP) phase 3, 5, and 6, including their common ancestor models. Models are distinguished by their complexity into atmosphere general circulation models (AGCMs), atmosphere–ocean GCMs (AOGCMs), and Earth system models (ESMs), indicated by color. 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 grey 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. Table S1 contains source data corresponding the this figure including literature references for the model relations.

laborating organizations). Publicly available or widely licensed models usually have much greater participation in CMIP and an outsized impact in the MMEs.

Relations between model code can often 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 implementations from a parent model and developing their own atmospheric dynamics. Likewise, the ocean, land, sea ice, and biochemistry components are swapped for other components in some derived models. 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 *ancestry weighting* and several simpler types of weighting impact the calculation of climate feedbacks and climate sensitivity in the CMIP MMEs. Zelinka et al. (2020) analyzed climate feedbacks, ECS, and ERF in CMIP5 and CMIP6. We perform the same analysis using their estimates of model quantities (Zelinka, 2022), but with different methods of weighting. Figure 3 shows results analogous to Figure 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 *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 *ancestry 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 *ancestry* 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 given the very large spread in model cloud feedbacks in the CMIP MMEs.

Interestingly, when we quantify the difference in feedback strength between the CMIP6 and CMIP5 MMEs (Figure 3c), we see that the *ancestry weighting* reduces the difference in cloud feedbacks between the two CMIP phases substantially. The magnitude difference is reduced from 77 to -26 $\mathrm{mWm^{-2}K^{-1}}$ for the total cloud feedback, from 145 to -68 $\mathrm{mWm^{-2}K^{-1}}$ for the shortwave (SW) cloud feedback, and from -70 to 41 $\mathrm{mWm^{-2}K^{-1}}$ for the longwave (LW) cloud feedback. However, the net and residual feedback magnitude difference is increased from 61 to -71 $\mathrm{mWm^{-2}K^{-1}}$ and from 3 to -33 $\mathrm{mWm^{-2}K^{-1}}$, respectively. We define the root mean square difference (RMSD) between CMIP6 and CMIP5 calculated across the elementary feedbacks (Planck, water vapor (WV), lapse rate (LR), albedo, SW cloud, LW cloud) as:

$$
\begin{aligned}
\mathrm{RMSD} &= \left(\frac{1}{n}\sum_{i=1}^{n}(\lambda_{i,\mathrm{CMIP6}}-\lambda_{i,\mathrm{CMIP5}})^2\right)^{1/2}, \\
n &= 6, \\
\lambda_i &= (\lambda_{\mathrm{Planck}}, \lambda_{\mathrm{WV}}, \lambda_{\mathrm{LR}}, \lambda_{\mathrm{albedo}}, \lambda_{\mathrm{SWcloud}}, \lambda_{\mathrm{LWcloud}})_i\ , \qquad (1)
\end{aligned}
$$

where $\lambda_i$ are means of individual feedbacks calculated from either CMIP5 ($\lambda_{i,\mathrm{CMIP5}}$) or CMIP6 ($\lambda_{i,\mathrm{CMIP6}}$). When the RMSD is calculated from the *ancestry weighted* feedback means compared with *simple* means, it is reduced by about 40% from 67 to 41 $\mathrm{mWm^{-2}K^{-1}}$. 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

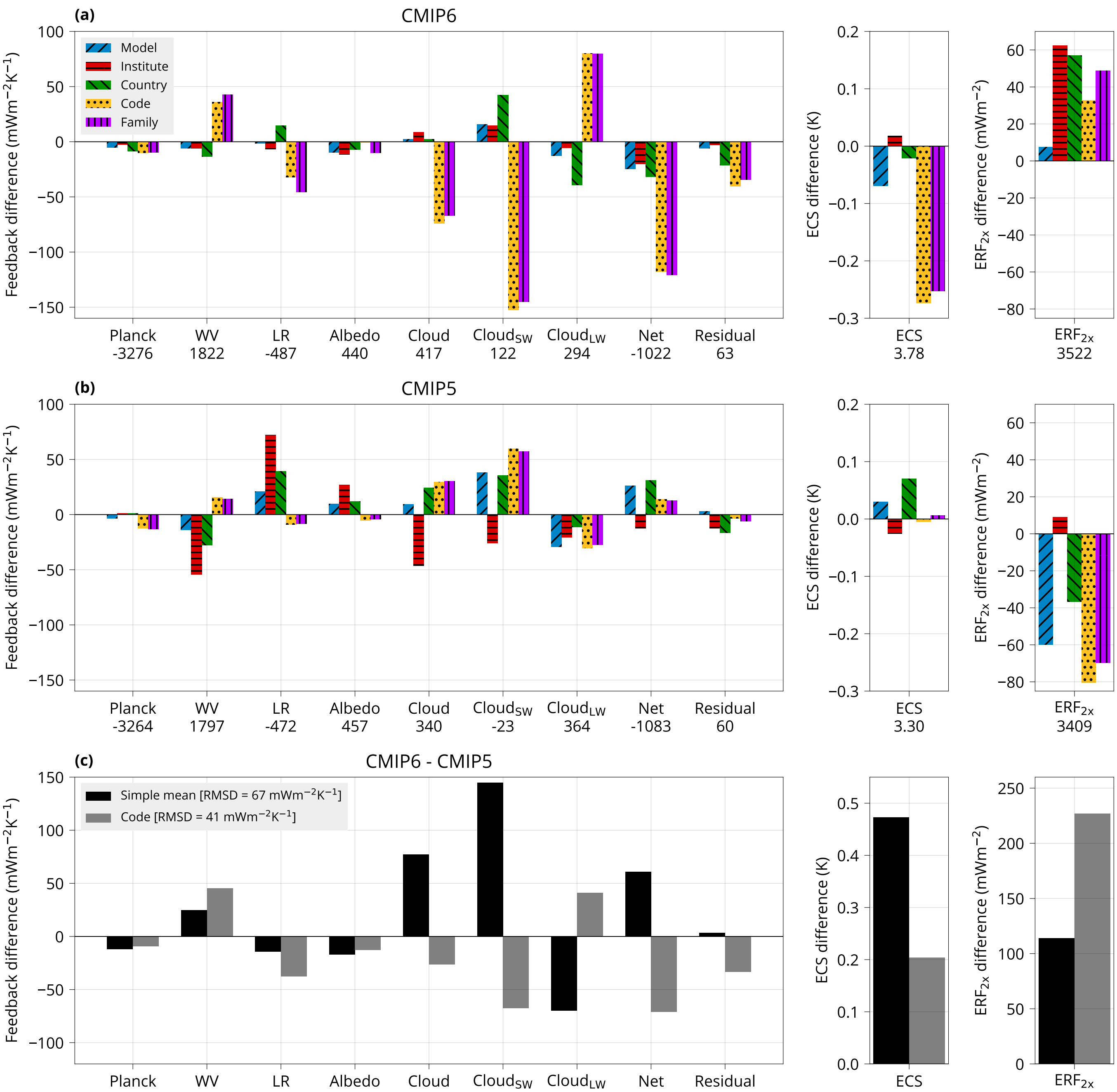


**Figure 3.** Climate feedbacks, effective climate sensitivity (ECS), and effective radiative forcing ($ERF_{2x}$) in the Coupled Model Intercomparison Project (CMIP) phases 6 **(a)** and 5 **(b)** under different weighting methods (*model*, *institute*, *country*, *ancestry*, and *family*) relative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

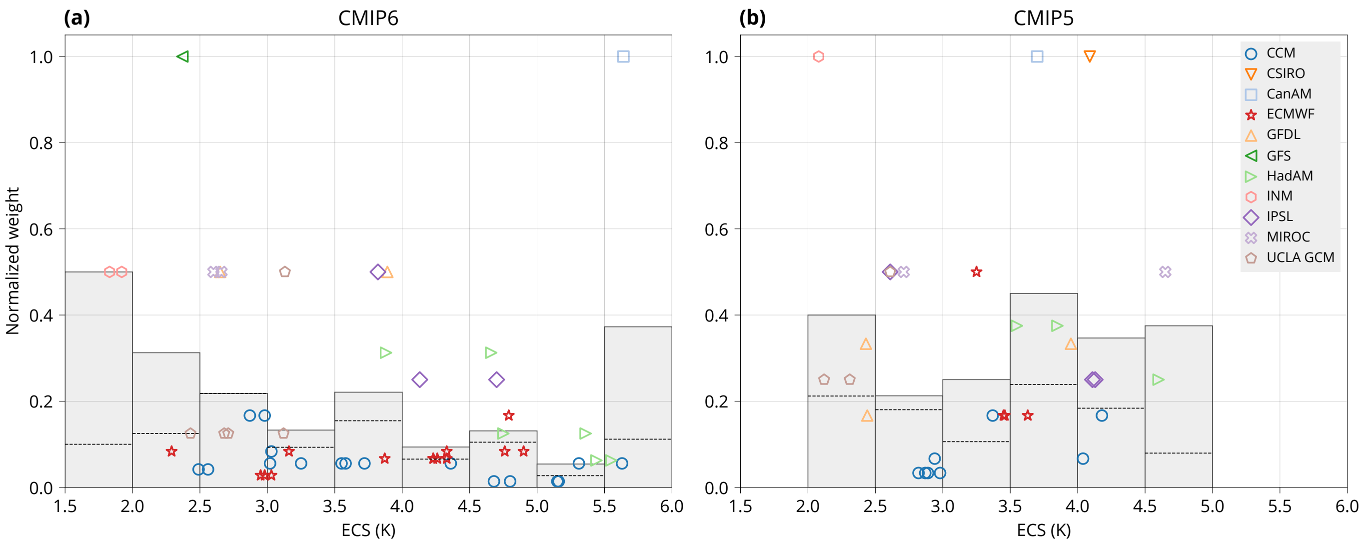


**Figure 4.** Statistical weights and effective climate sensitivity (ECS) of models in the Coupled Model Intercomparison Project (CMIP) phases 6 **(a)** and 5 **(b)** under the *ancestry weighting*. The model weights are normalized so that the maximum value is 1.0. The models are classified by their family, indicated by symbols. The shaded bars show a *simple* mean of model weights in the corresponding range of ECS. The dashed lines show the same as the bars, but multiplied by the number of models in the ECS range and normalized to sum to one.

account model code dependence. When the RMSD is calculated for *family weighting* (not shown in the plot), the RMSD is almost the same as *ancestry weighting* at 42 $\mathrm{mWm^{-2}K^{-1}}$. But it is less for the *model weighting* (reduced to 60 $\mathrm{mWm^{-2}K^{-1}}$), and a slight increase in RMSD is seen for *institute* (increased to 95 $\mathrm{mWm^{-2}K^{-1}}$) and *country* (increased to 79 $\mathrm{mWm^{-2}K^{-1}}$) weighting. This could mean that only the *ancestry*, *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 *ancestry 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 80 $\mathrm{mWm^{-2}}$ for $\mathrm{ERF_{2x}}$ in magnitude (Figure 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.). The difference in ensemble mean ECS between CMIP6 and CMIP5 becomes much smaller with *ancestry weighting*, falling from 0.47 K (*simple* mean) to 0.20 K (*ancestry weighting*), but the difference in $\mathrm{ERF_{2x}}$ is increased from 114 to 226 $\mathrm{mWm^{-2}}$. Thus, it is possible that a weighting method which accounts for model code dependency can explain some of the difference in ECS between CMIP5 and CMIP6 as resulting from an over-representation of models with high ECS in the CMIP6 ensemble.

Figure 4 shows model ECS and the statistical weights of models under the *ancestry weighting*. It can be seen that in CMIP6, the model weight is the highest for the lowest ECS range and progressively lower with increasing ECS (except for the highest ECS range), due to the fact that models with higher ECS are generally populated by the large model families HadAM, CCM, and to a lesser extent IPSL and ECMWF, while models with lower ECS come from more diverse families. Because of how the *ancestry weight-*

*ing* algorithm works, models in larger families generally have lower per-model weight. In CMIP5 model weights are more even across the ECS range than in CMIP6. Partly, the higher *simple* mean of ECS in CMIP6 is also the result of ECS above 5 K being populated by models, whereas in CMIP5 there are no models in this range. Thus, the higher *simple* mean ECS in CMIP6 can be attributed mostly to the HadGEM and CCM model families, and their effect is reduced under the *ancestry weighting* by smaller per-model weight given to models in large model families. Figure 4 also shows the weights multiplied by the number of models in each ECS range (dashed lines). While the two most extreme ECS ranges in CMIP6 (below 2 K and above 5.5 K) have relatively large per-model weights, the number of models in these ranges is small (two), and they have little overall effect on the *ancestry-weighted* ECS mean.

### 4.3 Climate Feedbacks and Sensitivity by Model Family

We analyzed climate feedbacks and sensitivity by model family (Figure 5). Because model *family weighting* showed results similar to *ancestry weighting* (section 4.2), it should be a good proxy for *ancestry 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 (in this section we refer to *simple* mean as "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 its models in CMIP6, but a small spread in CMIP5. Despite this, they have some characteristic properties, such as in mostly above-mean total and SW cloud feedback and below-mean LW cloud feedback in CMIP6; mostly below-mean total cloud feedback, but also above-mean lapse rate and surface albedo, and below-mean water vapor feedback in CMIP5. 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 families of models show a large and relatively even spread around the multi-model mean. In this case, the *ancestry* or *family* weighting is unlikely to make a significant difference in terms of the influence of the family on the overall MME 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. ECS in of the HadAM family is significantly above-mean, and ECS of the UCLA GCM family is significantly below-mean (at 95% confidence).

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 families have substantial interdependence 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 analyze the impact of the *ancestry* 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. Figures 6 and 7 show GMST time series in the CMIP6 and CMIP5 experiments (respectively), grouped by model family, as well as *family* and *ancestry weighted*

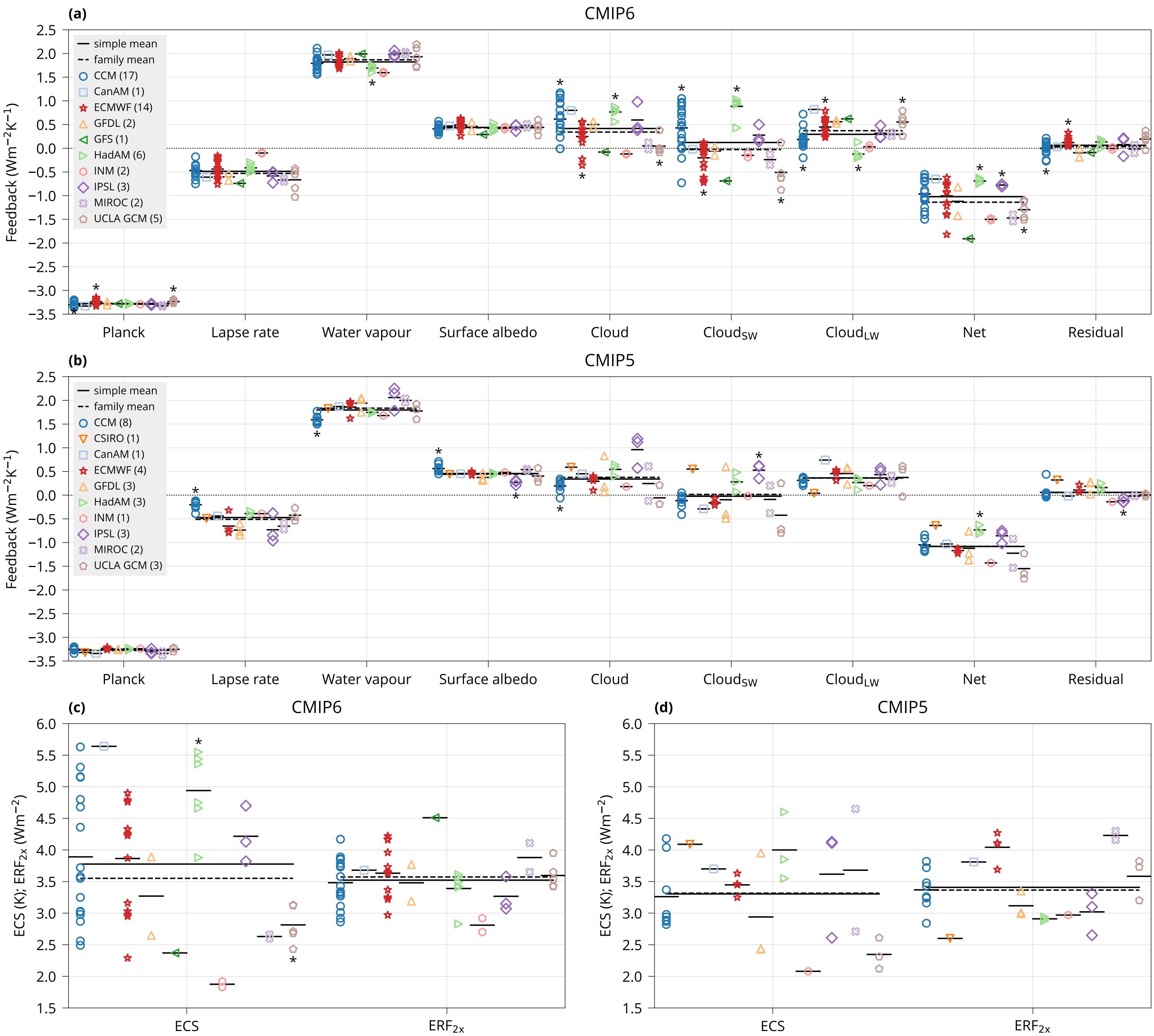


**Figure 5.** Climate feedbacks, effective climate sensitivity (ECS), and effective radiative forcing ($ERF_{2x}$) arranged by model family in the Coupled Model Intercomparison Project (CMIP) phases 5 **(b, d)** and 6 **(a, c)**. Model family is identified by the oldest ancestor model. In the legend, 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).

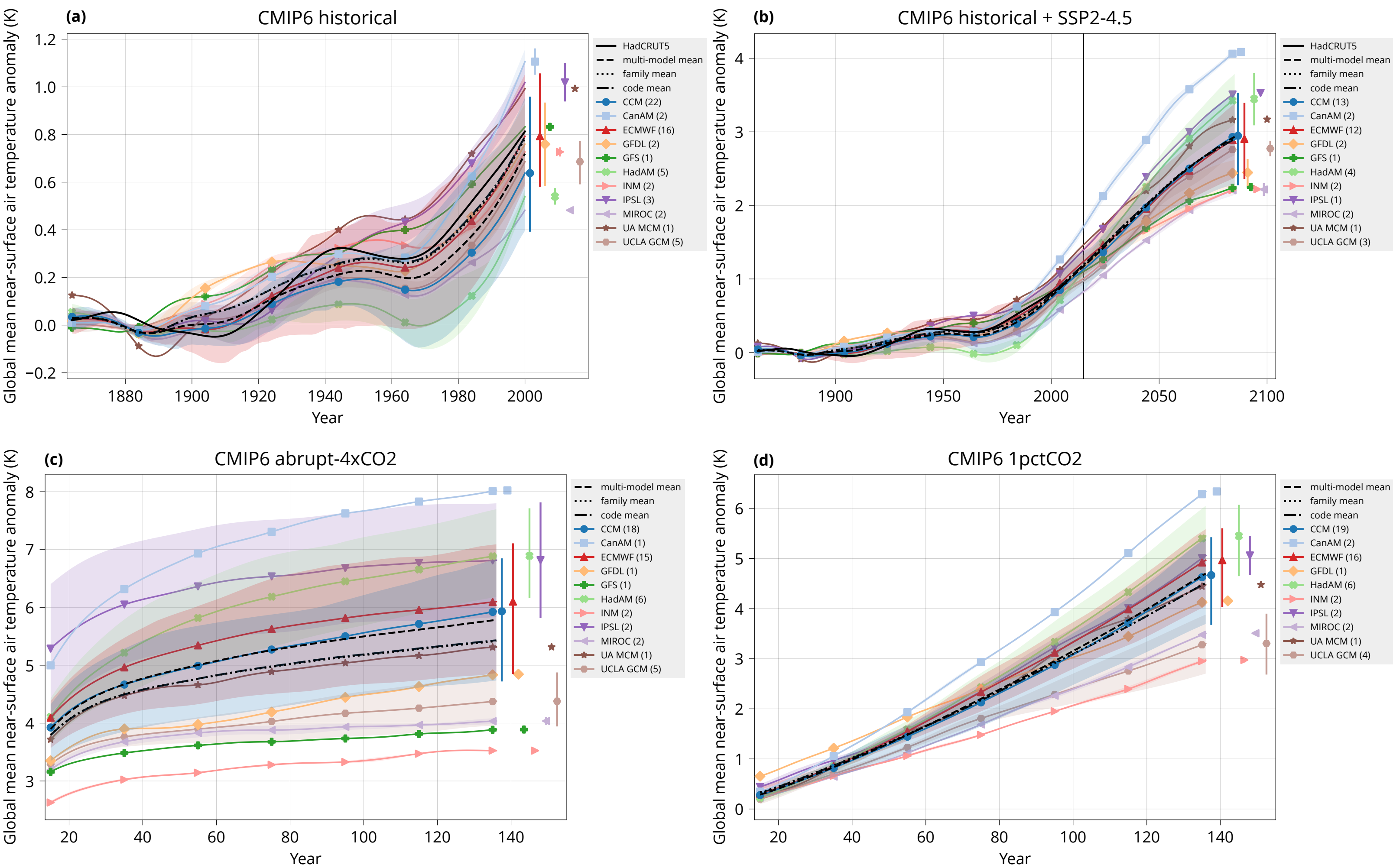


**Figure 6.** Time series of global mean near-surface temperature in CMIP6 experiments by model family and the *simple* multi-model, *ancestry*, and *family* mean (section 3.2). The model family time series are a *simple* mean of models in the family. The time series are smoothed with a Gaussian kernel with a standard deviation of 7 years. The first and the last 14 years of the time series are not shown to avoid artifacts caused by 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)**. Shaded areas are confidence bands representing the 68$^{\text{th}}$ percentile range. The vertical divider in the *historical* + SSP2-4.5 plot separates the time ranges 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.

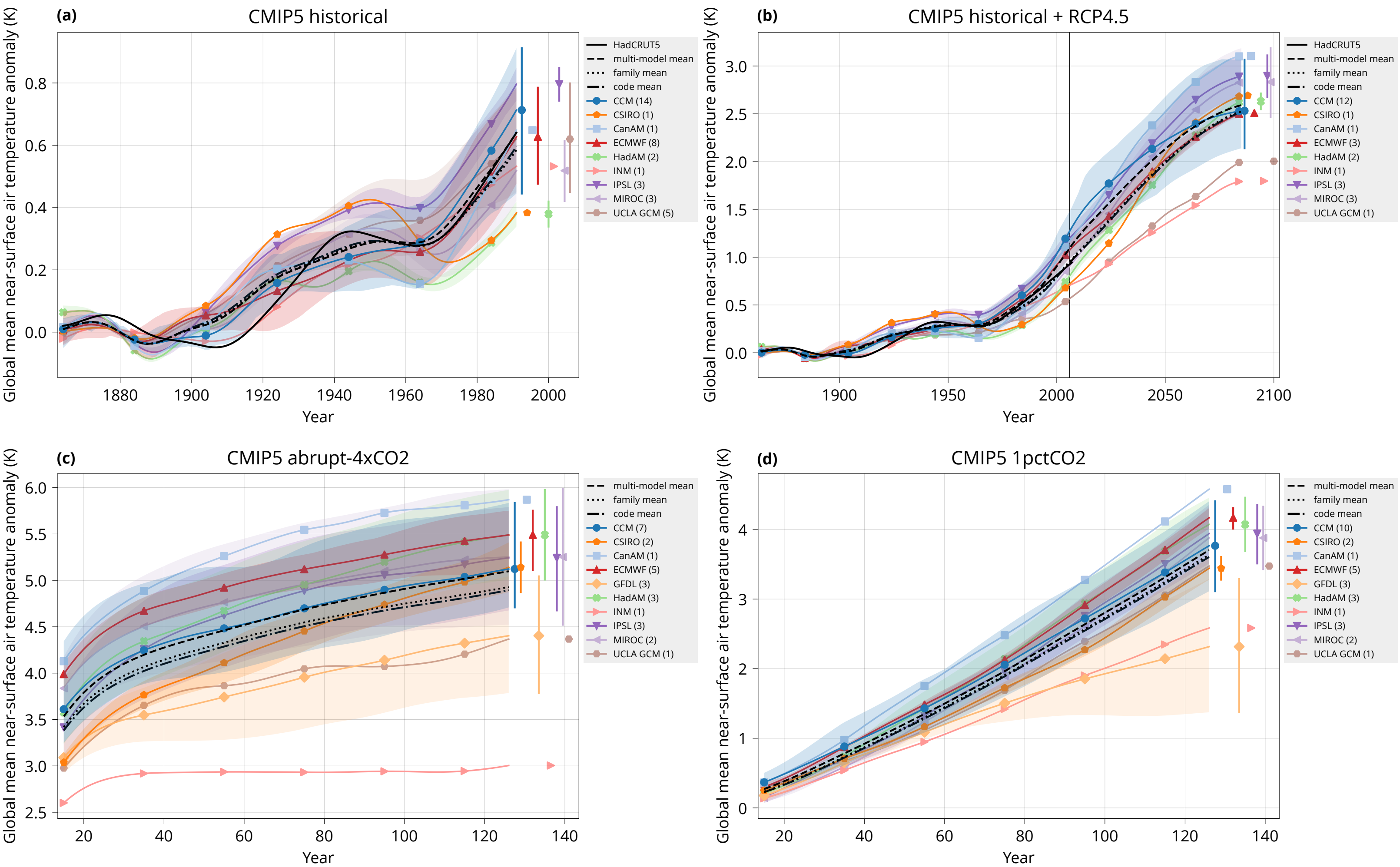


**Figure 7.** The same as Figure 6 but for CMIP5, and the RCP4.5 experiment instead of SSP2-4.5.

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, depending on the experiment and CMIP phase), 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 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 experiments *historical* + RCP4.5 (combined experiments), *abrupt-4xCO2*, and *1pctCO2* is relatively narrow. The other larger family HadAM has a relatively small spread in most experiments, consistent with the results of section 4.3. Notably, in the CMIP6 *historical* experiment, HadAM is 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 (each 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.

The *family* and *ancestry 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 *ancestry 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 also more consistent with observations, whereas in the *1pctCO2* and *abrupt-4xCO2* experiments it is colder than the *simple* mean (in both CMIP6 and CMIP5). When CMIP6 is compared with CMIP5, model families tend 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 being 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 large amounts of code may have been replaced in some models, 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, are particularly large. The CCM-derived models were extensively forked 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 dominate CMIP6, accounting for 70% of all model runs, an increase from about 50% represented by the three largest model families in CMIP3 and CMIP5. Based on the code genealogy, we developed an *ancestry weighting* method, the aim of which was to more fairly weigh code-related models than a *simple* multi-model mean, thus mitigating structural model dependence effects in MMEs. We showed that when applied on CMIP5 and CMIP6, the *ancestry* and *family weighting* produced substantial differences in the climate feedbacks, sensitivity, and forcing, especially the cloud feedbacks (total, shortwave and longwave), ECS, and $\mathrm{ERF_{2x}}$ relative to the difference in *simple* mean between CMIP6 and CMIP5 and relative to the standard deviation of the quantities in CMIP5 and CMIP6. The *ancestry* and *family weighting* methods produce very similar results. The *ancestry* and *family weighting* seem to be able to explain some of the difference between CMIP6 and CMIP5 (about 40% RMSD reduction in climate feedbacks, and about 60% RMSD reduction in ECS under the *ancestry 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 analyze climate feedbacks, sensitivity, and forcing by model family revealed that models in some families gave narrowly similar results (HadAM and UCLA GCM), and others in some cases had relatively wide spread but consistently above- or below-mean values (ECMWF and CSM). This suggests that code similarity in some cases translates to similarities in climate properties, but in other cases there is a large spread despite model similarity. Lastly, we analyzed 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 *ancestry 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 these methods can partially balance the effect of the over-representation of model families with multiple similar models, like HadAM. Model families tend to exhibit tendencies toward greater or lower warming than the MME mean in response to increased $CO_2$ across the CMIP generations.

A limitation of our method of weighting based on model families or model code genealogy is that we have not quantified model similarity in other ways than through inheritance. 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. For example, it is possible that models which are related in the genealogy diverged enough from their ancestral models that it would be warranted to classify them as a separate family. This means that some models can be unjustly underweighted because they are grouped together with models to which they do not bear much resemblance or were developed for a different purpose in mind (discussed below). Overcoming this limitation would be a relatively difficult task. While it might be possible to investigate individual schemes and components in models to partially quantify the statistical distances between related models, it would be difficult to do so objectively. Such information is also unlikely to be available for all the CMIP participating models. Another possibility would be to analyze the code of models to quantify their similarity. A method of accurately quantifying similarity would necessitate analyzing large code bases, distinguishing scientific calculations from technical code, accounting for the fact that small changes in code can produce large differences in model results, and accounting for model runtime configuration. Emerging methods of code analysis based on deep artificial neural networks (DANNs) have a potential to be used for this task. DANN-based tools such as OpenAI Codex (Chen et al., 2021; OpenAI, 2023), GitHub Copilot (GitHub, 2023) and DeepMind AlphaCode (DeepMind, 2023) have been developed to translate natural text to computer code. This approach has a potential to be adapted to quantifying code similarity. However, regardless of the availability of such methods, access to the model code would be necessary. This is a substantial hurdle given that most model code is closed-source. Apart from this, the source code of older models (dating back several decades) might not be readily available even to the current modeling groups, or even preserved at all. In summary, users of our model code genealogy should be mindful that the proposed weighting methods are only a "first-order" approximation of model similarity, and they should make an educated choice when selecting models for an analysis or deciding which models to include in a model family for the purpose of weighting.

Structural dependence between code-related models is sometimes reduced by diverging purposes of models. We did not make an attempt to quantify this because limitations similar to those mentioned above. The purpose of a model, such as a geographical, process, or quantity focus, is only rarely explicitly stated and it would be difficult to objectively quantify this divergence. In such case the *family* and *ancestry weighting* can give too little weight to those models in the same family or branch of the code genealogy which are substantially different from the rest of the models due to their purpose. One way in which models are divergent within the same family or branch is their complexity in terms of being an AGCM, AOGCM or ESM (Figure 2). It can be expected that ESMs are substantially different from a related AOGCM due to the inclusion of the carbon cycle, vegetation, atmospheric chemistry, biochemistry and other processes. Similarly AGCMs, even though rarely participating in CMIP as standalone models, are expected to differ substantially from related AOGCMs because they do not contain a prognostic ocean component. One way of accounting for this would be to analyze AOGCMs and ESMs separately. For example, Meehl et al. (2020) note that emissions feedbacks included in the ESM GFDL-ESM4 (Dunne et al., 2020) reduce ECS compared to its parent AOGCM GFDL-CM4 (Held et al., 2019); GFDL-ESM4 has ECS 3.9 K and GFDL-CM4 has ECS 2.6 K. In summary, the focus solely on model code inheritance as presented here does not account for this context, introducing limitations to our weighting methods.

To put our results into a broader perspective, we do not argue against the use of *simple* multi-model means, or model output and performance weighting methods in general, 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 interdependencies 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 easily implemented method of weighting for MME studies, especially if there is an expectation that model duplication is affecting the results.

The presented model code genealogy (Figure 2) can be further extended as more models become available in future CMIP phases. We provide the Scalable Vector Graphics (SVG) source of this figure so that it can be extended in the future, 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 have shown that as expected, code-related models tend to have related climate characteristics, which 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 have 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 ancestry weighting. This can be readily employed in MME assessments if a more fair model weighting is desired.

## Appendix A Model Ancestry Weight Calculation

Statistical weights in model *ancestry weighting* are calculated using the model code genealogy in Figure 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 Figure 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 Figure 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.

**Acknowledgments**

We thank the editor Tapio Schneider and two anonymous reviewers. We would like to acknowledge funding from the FORCeS project: "Constrained aerosol forcing for improved climate projections" (FORCeS project authors, 2023) and nextGEMS (nextGEMS project authors, 2023), funded by the European Union's Horizon 2020 research and innovation program under grant agreement numbers 821205 and 101003470, respectively, and funding from the Swedish e-Science Research Centre (SeRC). We acknowledge the World Climate Research Programme (WCRP), the Coupled Model Intercomparison Project (CMIP), the Earth System Grid Federation (ESGF), and the climate modeling groups for providing the model output data. We acknowledge the Met Office Hadley Centre for providing the HadCRUT5 dataset and Mark Zelinka for providing model climate feedback and climate sensitivity data. Last but not least, we thank the developers of the open source software Python, NumPy, Matplotlib, SciPy, Inkscape, and Devuan GNU/Linux, on which are work depended substantially.

## Open Research Section

Our data processing and visualization code, as well as the associated data are available publicly on GitHub (Kuma, 2022a) and Zenodo (Kuma, 2022b). The version used in our analysis is 1.0.0. The software is licensed under an open source license (MIT), the project internal data files and the output data files are in the public domain [Creative Commons license CC0, Creative Commons (2023b)], and the model code genealogy graph images and output plots are licensed under the Creative Commons Attribution 4.0 International license [CC BY 4.0, Creative Commons (2023a)]. CMIP5 and CMIP6 model output is publicly available on the Earth System Grid Federation websites (CMIP5, 2022; CMIP6, 2022). The input data for model ECS and climate feedbacks are available publicly (Zelinka, 2022). The HadCRUT5 data are available publicly (Met Office Hadley Centre, 2022). Our code was developed in Python version 3.9.2 (Python Software Foundation, 2023) on Devuan GNU/Linux version 4 (Devuan project authors, 2023). The following Python packages were used directly in our code: ds-format version 3.5.1, matplotlib version 3.7.1 (Hunter, 2007), numpy version 1.22.1 (Harris et al., 2020), pandas version 1.4.3 (Wes McKinney, 2010), pst version 2.0.0, pymc3 version 3.11.5 (Patil et al., 2010), and scipy version 1.7.3 (Virtanen et al., 2020), obtained from the Python Package Index (Python community, 2023). Figure 2 was made in Inkscape version 1.0.2 (Inkscape project authors, 2023). All of the listed software is available publicly under open source licenses.

## References


Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., ... Schmidt, G. A. (2019). Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, *10*(1), 91–105. Retrieved from https://esd.copernicus.org/articles/10/91/2019/ doi: 10.5194/esd-10-91-2019

Alexander, K., & Easterbrook, S. M. (2015). The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations. *Geoscientific Model Development*, *8*(4), 1221–1232. Retrieved from https://gmd.copernicus.org/articles/8/1221/2015/ doi: 10.5194/gmd-8-1221-2015

Arakawa, A. (2000). Chapter 1: A personal perspective on the early years of general circulation modeling at UCLA. In D. A. Randall (Ed.), *General circulation model development* (Vol. 70, pp. 1–65). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/S0074614200800492 doi: https://doi.org/10.1016/S0074-6142(00)80049-2

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Rashid, H., Uotila, P., ... Puri, K. (2013). The ACCESS coupled model: description, control climate and evaluation. *Australian Meteorological and Oceanographic Journal*, *63*(1), 41-64. doi: 10.1071/ES13004

Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. *Climate dynamics*, *41*(3), 885–900. doi: 10.1007/s00382-012-1610-y

Boé, J. (2018). Interdependency in multimodel climate projections: Component replication and result similarity. *Geophysical Research Letters*, *45*(6), 2771–2779. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL076829 doi: 10.1002/2017GL076829

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., & Sanderson, B. M. (2014). Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, *41*(5), 1803–1808. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL059205 doi: 10.1002/2014GL059205

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., ...

Zaremba, W. (2021). *Evaluating large language models trained on code.*

CMIP3. (2022). *WCRP Coupled Model Intercomparison Project phase 3 (CMIP3) [Dataset].* Retrieved from https://esgf-node.llnl.gov/projects/cmip3/ (last access: 1 August 2022)

CMIP5. (2022). *WCRP Coupled Model Intercomparison Project phase 5 (CMIP5) [Dataset].* Retrieved from https://esgf-node.llnl.gov/projects/cmip5/ (last access: 1 August 2022)

CMIP6. (2022). *WCRP Coupled Model Intercomparison Project phase 6 (CMIP6) [Dataset].* Retrieved from https://esgf-node.llnl.gov/projects/cmip6/ (last access: 1 August 2022)

Creative Commons. (2023a). *Attribution 4.0 International (CC BY 4.0).* Retrieved from https://creativecommons.org/licenses/by/4.0/ (last access: 11 May 2023)

Creative Commons. (2023b). *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication.* Retrieved from https://creativecommons.org/publicdomain/zero/1.0/ (last access: 11 May 2023)

DeepMind. (2023). *AlphaCode.* Retrieved from https://alphacode.deepmind.com (last access: 27 April 2023)

Devuan project authors. (2023). *Devuan GNU+Linux Free Operating System [Software].* Retrieved from https://www.devuan.org (last access: 11 May 2023)

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., ... Zhao, M. (2020). The GFDL Earth System Model version 4.1 (GFDL-ESM 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2019MS002015. (e2019MS002015 2019MS002015) doi: 10.1029/2019MS002015

Edwards, P. N. (2000a). *The agcm family tree.* Retrieved from http://pne.people.si.umich.edu/vastmachine/agcm.html (last access: 3 May 2023)

Edwards, P. N. (2000b). *Atmospheric general circulation modeling: A participatory history.* Retrieved from http://pne.people.si.umich.edu/sloan/mainpage.html (last access: 12 August 2022)

Edwards, P. N. (2000c). Chapter 2: A brief history of atmospheric general circulation modeling. In D. A. Randall (Ed.), *General circulation model development* (Vol. 70, pp. 67–90). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/S0074614200800509 doi: 10.1016/S0074-6142(00)80050-9

Edwards, P. N. (2011). History of climate modeling. *WIREs Climate Change*, *2*(1), 128–139. Retrieved from https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.95 doi: 10.1002/wcc.95

Edwards, P. N. (2013). Chapter 7: The infinite forecast. In *A vast machine: Computer models, climate data, and the politics of global warming* (pp. 139–186). The MIT Press.

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

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... Williamson, M. S. (2019, jan). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. Retrieved from https://doi.org/10.1038%2Fs41558-018-0355-y doi: 10.1038/s41558-018-0355-y

FORCeS project authors. (2023). *FORCeS: Constrained aerosol forcing for improved climate projections.* Retrieved from https://forces-project.eu (last access: 11 May 2023)

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D.,

… 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* (pp. 923–1054). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. doi: 10.1017/9781009157896.009

GitHub. (2023). *Copilot.* Retrieved from `https://github.com/features/copilot` (last access: 27 April 2023)

Gjermundsen, A., Nummelin, A., Olivié, D., Bentsen, M., Seland, Ø., & Schulz, M. (2021, Oct 01). Shutdown of Southern Ocean convection controls long-term greenhouse gas-induced warming. *Nature Geoscience*, *14*(10), 724-731. Retrieved from `https://doi.org/10.1038/s41561-021-00825-x` doi: 10.1038/s41561-021-00825-x

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., … Zhu, Q. (2019). The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2089–2129. doi: 10.1029/2018MS001603

Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S., … Taylor, K. E. (2013). Documenting climate models and their simulations. *Bulletin of the American Meteorological Society*, *94*(5), 623–627. Retrieved from `https://journals.ametsoc.org/view/journals/bams/94/5/bams-d-11-00035.1.xml` doi: 10.1175/BAMS-D-11-00035.1

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020, Sep). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from `https://doi.org/10.1038/s41586-020-2649-2` doi: 10.1038/s41586-020-2649-2

Haughton, N., Abramowitz, G., Pitman, A., & Phipps, S. J. (2015). Weighting climate model ensembles for mean and variance estimates. *Climate dynamics*, *45*(11), 3169–3181. doi: 10.1007/s00382-015-2531-3

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., … Zadeh, N. (2019). Structure and performance of gfdl's cm4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691-3727. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001829` doi: https://doi.org/10.1029/2019MS001829

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Inkscape project authors. (2023). *Inkscape: Draw freely [Software].* Retrieved from `https://inkscape.org` (last access: 11 May 2023)

Jebeile, J., & Crucifix, M. (2021). Value management and model pluralism in climate science. *Studies in History and Philosophy of Science*, *88*, 120-127. Retrieved from `https://www.sciencedirect.com/science/article/pii/S003936812100087X` doi: 10.1016/j.shpsa.2021.06.004

Jun, M., Knutti, R., & Nychka, D. W. (2008a). Spatial analysis to quantify numerical model bias and dependence. *Journal of the American Statistical Association*, *103*(483), 934–947. Retrieved from `https://doi.org/10.1198/016214507000001265` doi: 10.1198/016214507000001265

Jun, M., Knutti, R., & Nychka, D. W. (2008b). Local eigenvalue analysis of CMIP3 climate model errors. *Tellus A: Dynamic Meteorology and Oceanography*, *60*(5), 992–1000. Retrieved from `https://doi.org/10.1111/j.1600-0870.2008.00356.x` doi: 10.1111/j.1600-0870.2008.00356.x

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3), 395–404. doi: 10.1007/s10584-010-9800-2

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, *23*(10), 2739–2758. Retrieved from `https://journals.ametsoc.org/view/`

journals/clim/23/10/2009jcli3361.1.xml doi: 10.1175/2009JCLI3361.1
Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194–1199. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50256 doi: 10.1002/grl.50256
Krishnan, R., Swapna, P., Choudhury, A. D., Narayansetti, S., Prajeesh, A. G., Singh, M., ... Ingle, S. (2021). *The IITM Earth System Model (IITM ESM).* arXiv. doi: 10.48550/ARXIV.2101.03410
Kuma, P. (2022a). *Code accompanying the manuscript "Climate model code genealogy and its relation to climate feedbacks and sensitivity" (Version 1.0.0) [Software].* Retrieved from https://github.com/peterkuma/model-code-genealogy-2022/ (last access: 6 December 2022)
Kuma, P. (2022b). *Code accompanying the manuscript "Climate model code genealogy and its relation to climate feedbacks and sensitivity" (Version 1.0.0) [Software].* Zenodo. doi: 10.5281/zenodo.7407118
Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., & Seland, Ø. (2022). Machine learning of cloud types in satellite observations and climate models. *Atmospheric Chemistry and Physics*. (in press) doi: 10.5281/zenodo.7400969
Lenhard, J., & Winsberg, E. (2010). 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*(3), 253–262. doi: 10.1016/j.shpsb.2010.07.001
Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, *227*(7), 3431–3444. Retrieved from https://www.sciencedirect.com/science/article/pii/S0021999107000952 doi: 10.1016/j.jcp.2007.02.034
Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*(8). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011GL046864 doi: 10.1029/2011GL046864
Masson-Delmotte, V., et al. (Eds.). (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, Cambridge, United Kingdom.
Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., ... Taylor, K. E. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*(9), 1383–1394. Retrieved from https://journals.ametsoc.org/view/journals/bams/88/9/bams-88-9-1383.xml doi: 10.1175/BAMS-88-9-1383
Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., ... Schlund, M. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26), eaba1981. Retrieved from https://www.science.org/doi/abs/10.1126/sciadv.aba1981 doi: 10.1126/sciadv.aba1981
Mendlik, T., & Gobiet, A. (2016). Selecting climate simulations for impact studies based on multivariate patterns of climate change. *Climatic change*, *135*(3), 381–393. doi: 10.1007/s10584-015-1582-0
Met Office Hadley Centre. (2022). *HadCRUT5 [Dataset].* Retrieved from https://www.metoffice.gov.uk/hadobs/hadcrut5/ (last access: 12 December 2022)
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.
Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021). An updated assessment of near-surface temper-

ature change from 1850: The HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361 (e2019JD032361 2019JD032361) doi: 10.1029/2019JD032361

Morrison, M. A. (2021). *The models are alright: A socio-epistemic theory of the landscape of climate model development* (Unpublished doctoral dissertation). Indiana University, Indiana, United States.

nextGEMS project authors. (2023). *nextGEMS: Next Generation Earth Modelling Systems.* Retrieved from https://nextgems-h2020.eu (last access: 11 May 2023)

OpenAI. (2023). *Codex.* Retrieved from https://openai.com/blog/openai-codex (last access: 27 April 2023)

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457–477. doi: 10.1086/708691

Parker, W. S., & Winsberg, E. (2018, Jan 01). Values and evidence: how models make a difference. *European Journal for Philosophy of Science*, *8*(1), 125-142. Retrieved from https://doi.org/10.1007/s13194-017-0180-6 doi: 10.1007/s13194-017-0180-6

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, *35*(4), 1–81. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v035i04 doi: 10.18637/jss.v035.i04

Pennell, C., & Reichler, T. (2011). On the effective number of climate models. *Journal of Climate*, *24*(9), 2358–2367. Retrieved from https://journals.ametsoc.org/view/journals/clim/24/9/2010jcli3814.1.xml doi: 10.1175/2010JCLI3814.1

Pulkkinen, K., Undorf, S., Bender, F., Wikman-Svahn, P., Doblas-Reyes, F., Flynn, C., … Thompson, E. (2022, Jan 01). The value of values in climate science. *Nature Climate Change*, *12*(1), 4–6. Retrieved from https://doi.org/10.1038/s41558-021-01238-9 doi: 10.1038/s41558-021-01238-9

Pulkkinen, K., Undorf, S., & Bender, F. A.-M. (2022, Nov 18). Values in climate modelling: testing the practical applicability of the Moral Imagination ideal. *European Journal for Philosophy of Science*, *12*(4), 68. Retrieved from https://doi.org/10.1007/s13194-022-00488-4 doi: 10.1007/s13194-022-00488-4

Python community. (2023). *Python Package Index.* Retrieved from https://pypi.org (last access: 11 May 2023)

Python Software Foundation. (2023). *Python project [Software].* Retrieved from https://www.python.org (last access: 11 May 2023)

Remmers, J. O., Teuling, A. J., & Melsen, L. A. (2020). Can model structure families be inferred from model output? *Environmental Modelling & Software*, *133*, 104817. Retrieved from https://www.sciencedirect.com/science/article/pii/S1364815219308436 doi: 10.1016/j.envsoft.2020.104817

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016, apr). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. Retrieved from https://doi.org/10.7717/peerj-cs.55 doi: 10.7717/peerj-cs.55

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015a). Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate*, *28*(13), 5150–5170. Retrieved from https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00361.1.xml doi: 10.1175/JCLI-D-14-00361.1

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015b). A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, *28*(13), 5171–5194. Retrieved from https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00362.1.xml doi: 10.1175/

JCLI-D-14-00362.1

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B., Fisher, R. A., & Knutti, R. (2021). The potential for structural errors in emergent constraints. *Earth System Dynamics*, *12*(3), 899–918. Retrieved from `https://esd.copernicus.org/articles/12/899/2021/` doi: 10.5194/esd-12-899-2021

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., & Eyring, V. (2020). Emergent constraints on equilibrium climate sensitivity in CMIP5: do they hold for CMIP6? *Earth System Dynamics*, *11*(4), 1233–1258. Retrieved from `https://esd.copernicus.org/articles/11/1233/2020/` doi: 10.5194/esd-11-1233-2020

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., ... Saha, S. (2017). Practice and philosophy of climate model tuning across six us modeling centers. *Geoscientific Model Development*, *10*(9), 3207–3223. Retrieved from `https://gmd.copernicus.org/articles/10/3207/2017/` doi: 10.5194/gmd-10-3207-2017

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., ... Zelinka, M. D. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000678` (e2019RG000678 2019RG000678) doi: 10.1029/2019RG000678

Steinschneider, S., McCrary, R., Mearns, L. O., & Brown, C. (2015). The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning. *Geophysical Research Letters*, *42*(12), 5014-5044. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL064529` doi: 10.1002/2015GL064529

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. Retrieved from `https://journals.ametsoc.org/view/journals/bams/93/4/bams-d-11-00094.1.xml` doi: 10.1175/BAMS-D-11-00094.1

Touzé-Peiffer, L., Barberousse, A., & Le Treut, H. (2020). The Coupled Model Intercomparison Project: History, uses, and structural effects on climate research. *WIREs Climate Change*, *11*(4), e648. Retrieved from `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.648` doi: 10.1002/wcc.648

Undorf, S., Pulkkinen, K., Wikman-Svahn, P., & Bender, F. A.-M. (2022, Oct 03). How do value-judgements enter model-based assessments of climate sensitivity? *Climatic Change*, *174*(3), 19. Retrieved from `https://doi.org/10.1007/s10584-022-03435-7` doi: 10.1007/s10584-022-03435-7

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Voosen, P. (2022). 'Hot' climate models exaggerate Earth impacts. *Science (New York, NY)*, *376*(6594), 685–685. doi: 10.1126/science.adc9453

Wang, C., Soden, B. J., Yang, W., & Vecchi, G. A. (2021a). Compensation between cloud feedback and aerosol-cloud interaction in CMIP6 models. *Geophysical Research Letters*, *48*(4), e2020GL091024. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091024` (e2020GL091024 2020GL091024) doi: 10.1029/2020GL091024

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). doi: 10.25080/Majora-92bf1922-00a

Williams, J., Morgenstern, O., Varma, V., Behrens, E., Hayek, W., Oliver, H., ...

Frame, D. (2016). Development of the New Zealand Earth System Model: NZESM. *Weather and Climate*, *36*, 25–44. doi: 10.2307/26779386

Winsberg, E. (2012). Values and uncertainties in the predictions of global climate models. *Kennedy Institute of Ethics Journal*, *22*(2), 111–137. Retrieved from https://muse.jhu.edu/pub/1/article/484359 doi: 10.1353/ken.2012.0008

Zelinka, M. D. (2022). *GitHub repository mzelinka/cmip56_forcing_feedback_ecs [Dataset]*. Retrieved from https://github.com/mzelinka/cmip56_forcing_feedback_ecs (last access: 3 August 2022)

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782 (e2019GL085782 10.1029/2019GL085782) doi: 10.1029/2019GL085782