

Climate Model Code Genealogy and its Relation to Climate Feedbacks and Sensitivity

Peter Kuma¹, Frida A.-M. Bender¹, and Aiden R Jönsson¹

¹Department 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.

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

63 came with difficulties because such a multi-model data set was not designed to repre-
64 sent structural model uncertainty in an unbiased way (Abramowitz et al., 2019). The
65 two most recent CMIP phases are phase 5 (Taylor et al., 2012) and phase 6 (Eyring et
66 al., 2016, 2019).

67 Modern climate models such as GCMs and ESMs are highly complex software, con-
68 sisting of many components, modules, and configuration parameters. Usually, compo-
69 nents such as the atmosphere, ocean, land, sea ice, chemistry, biology, and others are cou-
70 pled together continuously during a simulation (Alexander & Easterbrook, 2015). These
71 components may be divided into subcomponents, modules or schemes representing var-
72 ious physical parametrizations, such as radiative transfer in the atmospheric component.
73 Components and subcomponents can sometimes be easily replaced with others, or they
74 can be turned on or off depending on the configuration. These model parts have been
75 shared relatively freely between different models in the same modeling group as well as
76 between groups internationally (in the following text we will use the terms “modeling
77 group” and “institute”, the latter being common in the context of CMIP, interchange-
78 ably). Alexander and Easterbrook (2015) directly analyzed the source code of model com-
79 ponents, showing significant sharing of components between models thanks to their highly
80 modular nature. Furthermore, parametrizations documented in literature were imple-
81 mented in a variety of models, meaning that they use many of the same parametriza-
82 tions for certain physical processes. This development approach leads to structural model
83 dependence, which could mean that their model output is more similar than what would
84 be expected from structurally independent models. Understanding model structural de-
85 pendence is further complicated by the fact that only few models have publicly avail-
86 able source code. The practice of “forking” code, when a new branch of a code base is
87 created under a new name, is common in software development. This is also the case with
88 climate models, where different modeling groups base their work on forking of an exist-
89 ing model from the same or a different modeling group. This process can be quite opaque
90 to the end-users, who might, without access to further context, assume that a different
91 model name implies that the model is entirely independent. We can expect that model
92 code bases which are open source (such as the Community Earth System Model [CESM])
93 or licensed widely within international consortia (such as the Integrated Forecasting Sys-
94 tem [IFS]/ARPEGE and Hadley Centre Global Environmental Model [HadGEM]) are
95 more highly represented in model ensembles due to the ease of sharing code (Sanderson
96 et al., 2015b). This is potentially in contrast to the proliferation of code which produces
97 the best results, which could otherwise arise if all model code were openly available. As
98 discussed below, what constitutes “the best results” may be difficult to quantify and is
99 not guaranteed to coincide with the best projections. Guilyardi et al. (2013) initiated
100 better model and experiment metadata collection within CMIP5 in order to provide per-
101 tinent information to those performing research based on model comparisons.

102 Because all models are imperfect representations of reality, they are affected by var-
103 ious uncertainties in the model output, which can be broadly categorized as data, pa-
104 rameter, and structural uncertainty (Remmers et al., 2020). While data and parameter
105 uncertainty can be relatively easily quantified and sampled, structural uncertainty per-
106 taining to model code is hard to quantify or sample, and some authors noted that struc-
107 tural uncertainty is insufficiently sampled in CMIP MMEs (Knutti et al., 2010). Mod-
108 els participating in CMIP are dependent in a number of ways, including being essentially
109 the same model with a different configuration, sharing parts of their codes, model com-
110 ponents, and schemes, using the same data sets for validation, and implementing sim-
111 ilar parametrizations. Some authors have therefore called this MME an “ensemble of op-
112 portunity” (Masson & Knutti, 2011; Knutti et al., 2013; Sanderson et al., 2015a; Boé,
113 2018), since the inclusion is based on the intent of a modeling group to participate rather
114 than objective selection criteria. If model dependence is not taken into account, the cal-
115 culation of means, variance, and uncertainty can be biased, and spurious correlations (such
116 as in emergent constraints) can arise in an MME (Caldwell et al., 2014; Sanderson et al.,

117 2021). Remmers et al. (2020) investigated whether model code genealogy can be inferred
 118 from model output [also investigated earlier by Knutti et al. (2013) and discussed be-
 119 low]. Using a modular modeling framework, they generated a model ensemble of hydro-
 120 logical models by sampling the model “hypothesis space” [as defined in Remmers et al.
 121 (2020)] and compared its genealogies based on model code and model output. They found
 122 that it was not possible to infer complete model code genealogy based on model output
 123 because the performance of the inference was low. It is possible that the same would par-
 124 tially apply to much more complex models like GCMs and ESMs, and model code re-
 125 lationship needs to be studied in order to sample the model hypothesis space. Pennell
 126 and Reichler (2011) tried to quantify the effective number of models in an MME of 24
 127 CMIP3 models based on model output error similarity, and found this to be about 8. In-
 128 creasing the number of ensemble models did not substantially increase the effective num-
 129 ber of models. Sanderson et al. (2015b) reached a similar conclusion, and found that the
 130 number of independent models calculated based on the model output in CMIP5 is much
 131 smaller than the total.

132 The simplest approach to analyzing an MME is “model democracy”, where each
 133 model is given an equal weight in statistical calculations. More sophisticated approaches
 134 proposed to address model dependence include weighting or selecting models. Selecting
 135 models can be regarded as an extreme form of weighting. Often suggested weighting meth-
 136 ods are based on model performance (“model meritocracy”), model output or code de-
 137 pendence, and diversity. The topic of climate model dependence and genealogy has been
 138 covered in many previous studies, most of which used the dependence of the model out-
 139 put (Jun et al., 2008a, 2008b; Masson & Knutti, 2011; Knutti et al., 2013; Bishop & Abramowitz,
 140 2013; Sanderson et al., 2015a; Haughton et al., 2015; Mendlik & Gobiet, 2016), while a
 141 focus on code dependence has been relatively rare (Alexander & Easterbrook, 2015; Stein-
 142 schneider et al., 2015). Boé (2018) distinguishes these two approaches as “a posteriori”
 143 and “a priori”. Knutti et al. (2013) developed a CMIP5 model genealogy based on a hi-
 144 erarchical clustering of model output. They found that models from the same institute
 145 were much closer in their model output than other models, and contemplated that out-
 146 put similarity could be used for model weighting or selection to eliminate biases due to
 147 near duplicate models. A more simple approach is “institutional democracy”, where one
 148 model per modeling group is selected, and “component democracy”, where models are
 149 selected to represent different model components (Abramowitz et al., 2019). Edwards
 150 (2000a, 2000b, 2000c, 2011, 2013) described the early to modern history of climate mod-
 151 eling and constructed a partial “family tree” of atmospheric GCMs based on their code
 152 heritage. Another account on early climate modeling was given by Arakawa (2000). Boé
 153 (2018) summarized institute, atmospheric, oceanic, land, and sea ice components of CMIP5
 154 models and how they relate to proximity of the model results. However, the code depen-
 155 dence of all CMIP3, CMIP5, and CMIP6 models has not been analyzed. Partially, such
 156 understanding is limited by the availability of the source code. This contributes to the
 157 treatment of models as “black boxes” by the research community. Haughton et al. (2015)
 158 compared simple weighting with model performance and model output dependence weight-
 159 ing. They found performance weighting improved mean relative to observations (as ex-
 160 pected) but degraded variance estimation, and dependence weighting improved both. Steinschneider
 161 et al. (2015) identified close correlations between model output of models of the same
 162 family even on a regional scale, and showed that the clustering of similar models can re-
 163 sult in narrowing the MME variance attributable to intermodel correlations.

164 Reducing the size of an MME to a set of independent models is a relatively sim-
 165 ple method of avoiding model dependence. Sanderson et al. (2015b) noted that permit-
 166 ting only one model per institute in an MME could lead to unfairly dismissing models
 167 which are substantially different, and overestimating independence in cases where code
 168 is shared between institutes. Weighting models by country can have some merit due to
 169 the fact that models are sometimes developed with a focus on accuracy over the region
 170 where the institute is located, and a model might be more extensively validated against

171 data from observations in the region. For example, the New Zealand Earth System Model
172 (NZESM) (in practice developed alongside HadGEM/UKESM) was developed to reduce
173 Southern Ocean biases (Williams et al., 2016); the Indian Institute of Tropical Meteorology
174 ESM (IITM ESM) has a special focus on the South Asian monsoon (Krishnan et
175 al., 2021); the Australian Community Climate and Earth System Simulator coupled model
176 (ACCESS-CM) has a focus on reducing uncertainties over the Australian region (Bi et
177 al., 2013); and the Energy Exascale Earth System Model (E3SM) aims to support the
178 U.S. energy sector decisions (Golaz et al., 2019). Weighting models by errors relative to
179 observations (performance weighting) is complicated by the fact that there can be a de-
180 coupling between a climate model’s accuracy in representing present-day and historical
181 climate variables and its accuracy in representing the projected change (or trend) of the
182 variables under a climate scenario (Jun et al., 2008a; Zelinka, 2022; Kuma et al., 2022).
183 Thus, a model’s performance in future climate projections cannot be fully inferred from
184 its performance in present-day and historical climate. Performance weighting can also
185 favor models which are better tuned to present-day, historical or paleontological obser-
186 vations by compensating biases. It is possible that model quality cannot be estimated
187 solely from model output due to the fact that some models might represent physics more
188 consistently with our knowledge of fundamental physics, yet give inferior output when
189 compared to observations if they have fewer compensating biases or are tuned less to rep-
190 resent present-day or historical observations. Knutti (2010) provides a high-level discus-
191 sion of the topic of model democracy, uncertainty, weighting, evaluation, calibration and
192 tuning in the context of decision making.

193 Apart from explicit model weighting or selection choices, seldomly recognized implicit
194 choices based on values (other than widely acknowledged epistemic values such as
195 openness, objectivity, evidence, and impartiality) influence model development, evalu-
196 ation, selection, weighting, interpretation, and communication of results (Pulkkinen, Un-
197 dorf, Bender, Wikman-Svahn, et al., 2022; Pulkkinen, Undorf, & Bender, 2022; Lenhard
198 & Winsberg, 2010; Winsberg, 2012; Undorf et al., 2022). The climate system is too com-
199 plex to be captured by models perfectly. Some of the limitations stem from limited com-
200 putational resources, uncertainty about how to represent processes at a coarse level through
201 parametrizations, and a lack of observational data. Thus, model construction necessi-
202 tates and is affected by decisions regarding a variety of compromises. Traditionally, a
203 pursuit of purely knowledge-oriented science has been desired in order to avoid conclu-
204 sions distorted by scientists’ views, values and interests. However, some authors empha-
205 size that purely knowledge-oriented construction of climate models is impossible because
206 of decisions involved in the model development (Parker & Winsberg, 2018; Parker, 2020;
207 Jebeile & Crucifix, 2021; Morrison, 2021). These decisions can be driven by not only the
208 desire for creating an unbiased objective representation of the climate system, but also
209 by purposes, views, values, interests and limitations. They include for example a spe-
210 cific focus on modeling a certain geographical region and quantities of interest, the avail-
211 ability of validation data influenced by locations of observations, compromises regard-
212 ing what errors are permissible, types of tuning (Schmidt et al., 2017), decisions involved
213 in earlier versions of the same model or ancestral models resulting in inherited values,
214 limited knowledge and time of the researchers, and limited resources. In turn, they can
215 also perpetuate certain types of societal biases against traditionally understudied and
216 underrepresented regions. Rarely are such decisions or values and interests which drive
217 them explicitly acknowledged, which makes it difficult to quantify their impact on MMEs.
218 Although less acknowledged, interests can also include reasons for pursuing certain re-
219 search or development which are not driven by practical reasons but by curiosity. In a
220 broader view, the development of climate models has aspects of iterative development,
221 inheritance, recombination, cooperation, competition and filling of different niches. In
222 this way, it can be considered a collective optimization process with the goal of describ-
223 ing the important and diverse properties of the climate system (as considered by var-
224 ious actors) through pluralism in the face of limited knowledge and computational re-
225 sources, both of which also keep changing.

226 We can define the structure (code) of a model as based on a set of hypotheses about
 227 reality as well as computational realizations of such hypotheses. A desirable feature of
 228 an MME would be that models represent samples from the hypothesis space with prob-
 229 ability equal to our degree of belief that the hypothesis is true (note that this is differ-
 230 ent from a uniform sampling of the hypothesis space, which would be both impossible
 231 and undesirable due to its size). However, this is rarely the case with existing MMEs,
 232 and it is not easily quantifiable. It is generally not desirable that the model output of
 233 individual models in an MME is the most unique, because one would still want all mod-
 234 els to converge as closely as possible on the true representation of physical processes. Here,
 235 we define a “true representation” in limited terms as a pragmatically-oriented concep-
 236 tualization of the Earth system, which for example might not include the anthroposphere
 237 as commonly externalized in CMIP models through scenarios. Models can be similar in
 238 their output because they are convergent on the best representation of reality or because
 239 of code similarity, and this limits the use of model output as a measure of model depen-
 240 dence. We note that some authors advocate against a value-free ideal to which models
 241 should converge (Parker & Winsberg, 2018; Parker, 2020).

242 As a conceptual model (Figure 1), we can consider models in an MME to be sam-
 243 ples corresponding to representations of a physical reality in a hypothesis space. Here,
 244 representation is supposed to mean code which produces output for given initial and bound-
 245 ary conditions, i.e. without considering internal variability. While the true physical rep-
 246 resentation is unknown and impossible to simulate due to computational constraints, our
 247 collective belief that a given representation is true can be conceptualized theoretically
 248 by a probability density function (PDF). Ideally, models in an MME are independent
 249 samples from this PDF (Figure 1a). In actual MMEs (Figure 1b), however, models are
 250 dependent and tend to be clustered together for reasons incompatible with the PDF, such
 251 as the inclusion of several configurations or resolutions of a single model, selective shar-
 252 ing of code between models for reasons other than meritocracy (such as availability or
 253 political and organizational decisions), or model output availability. Therefore, if a PDF
 254 or its statistics are estimated from this MME, they will be biased compared to the ac-
 255 tual PDF. The aim is then to compensate for this bias with appropriate model weight-
 256 ing, selection or more sophisticated techniques such as emergent constraints. Even if we
 257 could estimate the PDF in an unbiased way, the value with the maximum likelihood or
 258 the mean are unlikely to coincide with the true physical representation, because such a
 259 PDF only represents our belief that a given physical representation is true, which is lim-
 260 ited by our knowledge. Note that model dependence itself does not preclude that an es-
 261 timate of the PDF is unbiased. For example, in the Metropolis algorithm (Metropolis
 262 et al., 1953), an unbiased estimate of a PDF is generated by sequentially producing a
 263 chain of samples which are close to each other. After a large enough number of itera-
 264 tions, an unbiased estimate of the PDF can be inferred from the collection of all sam-
 265 ples, despite close correlation between adjacent samples in the chain. Other aspects not
 266 considered in Figure 1 are that our knowledge about the climate system is shaped by var-
 267 ious decisions such as which parts of the climate system have been considered interest-
 268 ing to study or observe, and individual models are also affected by such decisions dur-
 269 ing their development. As mentioned above, some models even have a particular explic-
 270 itly stated purpose, such as ACCESS-CM, E3SM, IITM ESM and NZESM. The conse-
 271 quence of this is that models are not only biased samples of the PDF due to code de-
 272 pendence, but also due to value and interest-based decisions. For the same reasons they
 273 can also converge or diverge.

274 None of the model weighting methods mentioned above are without issues. Per-
 275 formance weighting can disregard models whose physics representation is relatively far
 276 from the most likely representation but still plausible, thus artificially narrowing the spread.
 277 Model dependence weighting based on output or code can disregard models which are
 278 close to other models but were chosen to be based on this model because of its perceived
 279 quality, thus preventing such an MME from narrowing down on the true representation

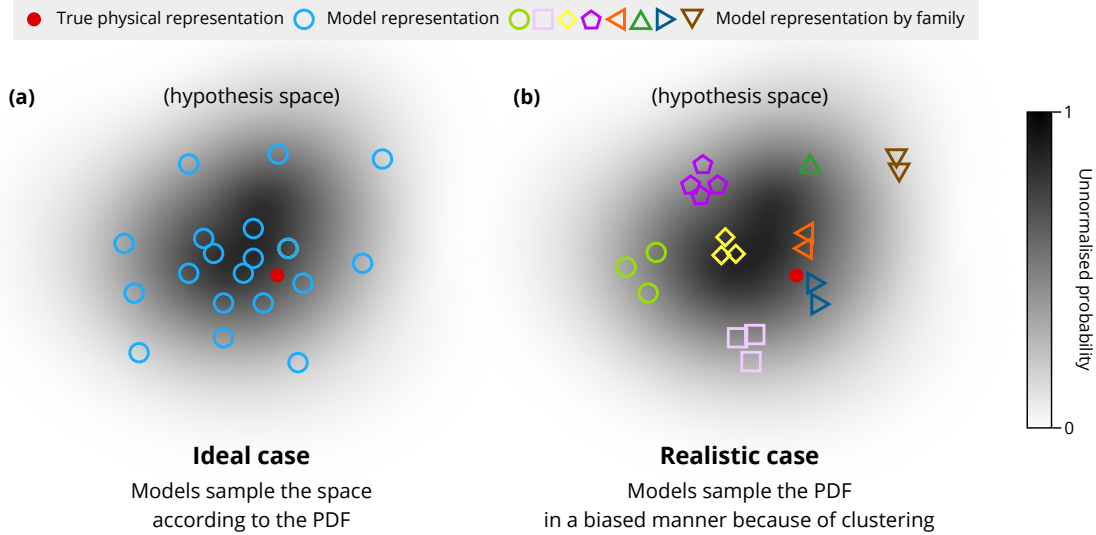


Figure 1. A theoretical illustrative example of model sampling of the model hypothesis space (model structural uncertainty), representing realizations of physical climate processes (model structure). The shading indicates a probability density function (PDF) quantifying our collective belief that a certain 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 (*pi-Control*), *historical*, Shared Socioeconomic Pathway 2-4.5 (*ssp245*), Representative Concentration Pathway 4.5 (*rcp45*), abrupt quadrupling of CO₂ (*abrupt-4xCO2*), and 1% yr⁻¹ CO₂ 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

330 global surface temperature dataset version 5 (HadCRUT5) (Morice et al., 2021) obtained
 331 from the Met Office Hadley Centre (2022). In order to analyze model code genealogy,
 332 we performed a broad literature survey, complemented by CMIP model metadata and
 333 information available online, particularly modeling groups’ websites. In total, we traced
 334 the genealogy of 167 models, of which 114 were participating in CMIP, and the rest were
 335 related to the CMIP models and thus necessary for reconstructing the genealogy. The
 336 model genealogy information, including related references, is also available in Table S1.
 337 Along with relations between models, we identified the model institute, the country where
 338 the institute resides, and the model family (defined by the oldest ancestral model in the
 339 genealogy). Model parameters such as ECS, TCR, ERF, and climate feedbacks were sourced
 340 from Zelinka et al. (2020) and the AR6. We use effective climate sensitivity calculated
 341 by Zelinka (2022), as an approximation of equilibrium climate sensitivity.

342 **3.2 Weighting Methods**

343 We applied several statistical weighting methods on the CMIP MMEs:

- 344 1. *Simple weighting.* Every model run is given equal weight. By “model run” we mean
 345 a model resolution or configuration (as listed in Table S1 in the columns *CMIP3/5/6*
 346 *names*), not multiple simulations performed with the same model but different ini-
 347 tial conditions.
- 348 2. *Family weighting.* Model families, defined as a complete branch as shown in Fig-
 349 ure 2 (discussed later in section 4.1), were given equal weight. This weight was
 350 further subdivided equally between models within the family.
- 351 3. *Institute weighting.* Model institutes, as shown in Figure 2 as labels on grey ar-
 352 eas, were given equal weight. This weight was further subdivided equally between
 353 models within the institute.
- 354 4. *Country weighting.* Model host countries, as shown in Figure 2 as labels on grey
 355 areas, were given equal weight. This weight was further subdivided equally be-
 356 tween models of the same country.
- 357 5. *Ancestry weighting.* The oldest ancestor models (marked with a thick outline in
 358 Figure 2) were given equal weight. This weight was subdivided gradually through
 359 branches to descendant models. This method is described in detail in Appendix
 360 Appendix A.
- 361 6. *Model weighting.* All models are given the same weight. This is different from the
 362 *simple weighting* – see the note below.

363 Note that in all of the above, if a model supplied multiple runs of different configura-
 364 tion or resolution, the model weight was further subdivided equally between the runs.
 365 For clarity, in the following text references to the weighting methods and weighted means
 366 corresponding to the methods above are *italicized*.

367 **3.3 Statistical Significance**

368 Statistical significance in climate feedbacks, sensitivity, and forcing in section 4.3
 369 was calculated using a Bayesian simulation with PyMC3 (Salvatier et al., 2016). The dif-
 370 ference between a *simple* mean of models within a family and a *simple* multi-model mean
 371 was marked as significant if the magnitude difference between the two means was larger
 372 than zero with 95% probability. The PyMC3 model is provided in the supplementary
 373 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 sub-derivatives 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 numerous 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-

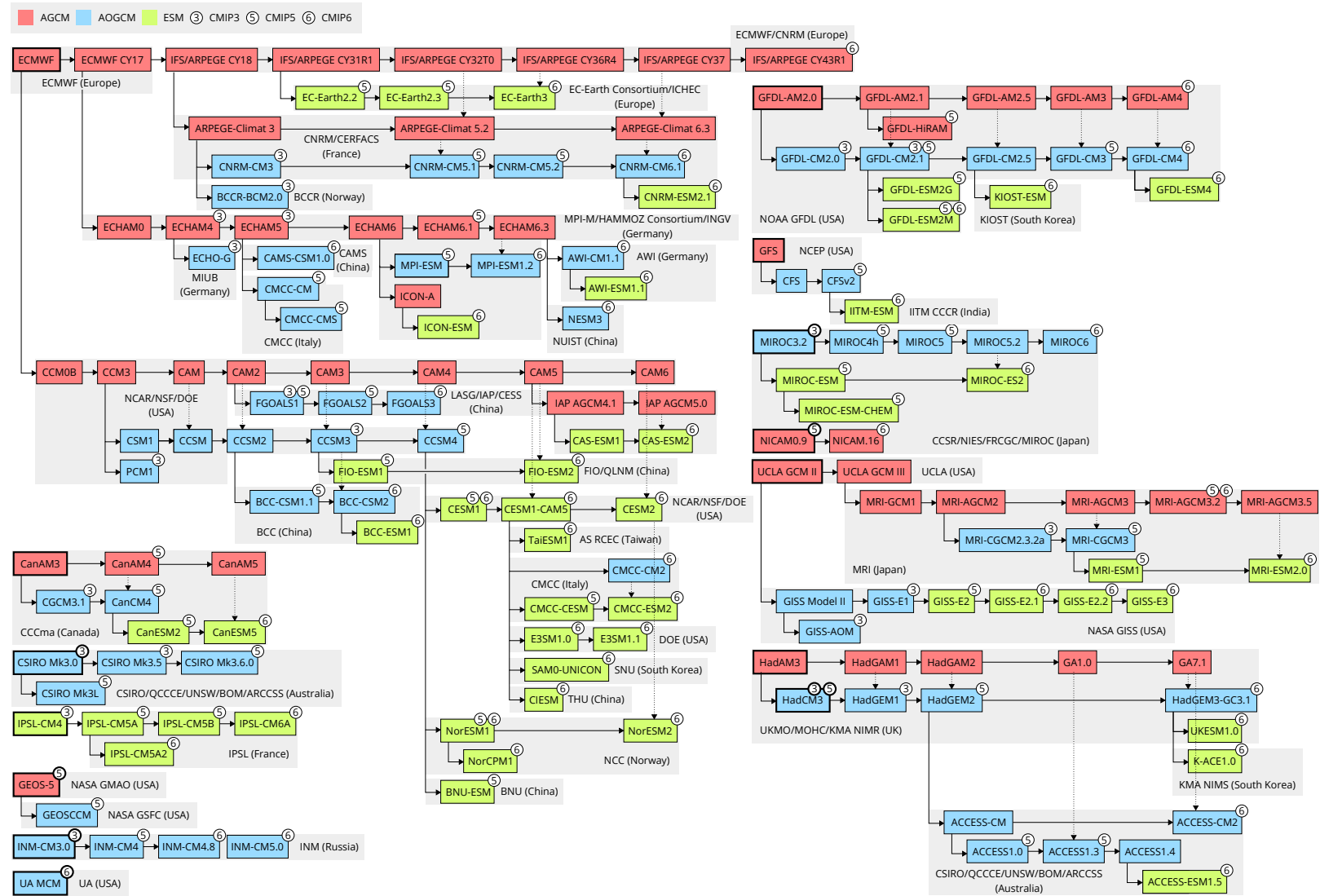


Figure 2. Model code genealogy of models participating in the Coupled Model Intercomparison Project (CMIP) phase 3, 5, and 6, including their common ancestor models. Models are distinguished by their complexity into atmosphere general circulation models (AGCMs), atmosphere–ocean GCMs (AOGCMs), and Earth 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 to 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 $150 \text{ mWm}^{-2}\text{K}^{-1}$ in magnitude in CMIP6 and $80 \text{ mWm}^{-2}\text{K}^{-1}$ 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 $-26 \text{ mWm}^{-2}\text{K}^{-1}$ for the total cloud feedback, from 145 to $-68 \text{ mWm}^{-2}\text{K}^{-1}$ for the shortwave (SW) cloud feedback, and from -70 to $41 \text{ mWm}^{-2}\text{K}^{-1}$ for the longwave (LW) cloud feedback. However, the net and residual feedback magnitude difference is increased from 61 to $-71 \text{ mWm}^{-2}\text{K}^{-1}$ and from 3 to $-33 \text{ mWm}^{-2}\text{K}^{-1}$, 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:

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

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 *ancestry weighted* feedback means compared with *simple* means, it is reduced by about 40% from 67 to $41 \text{ mWm}^{-2}\text{K}^{-1}$. 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

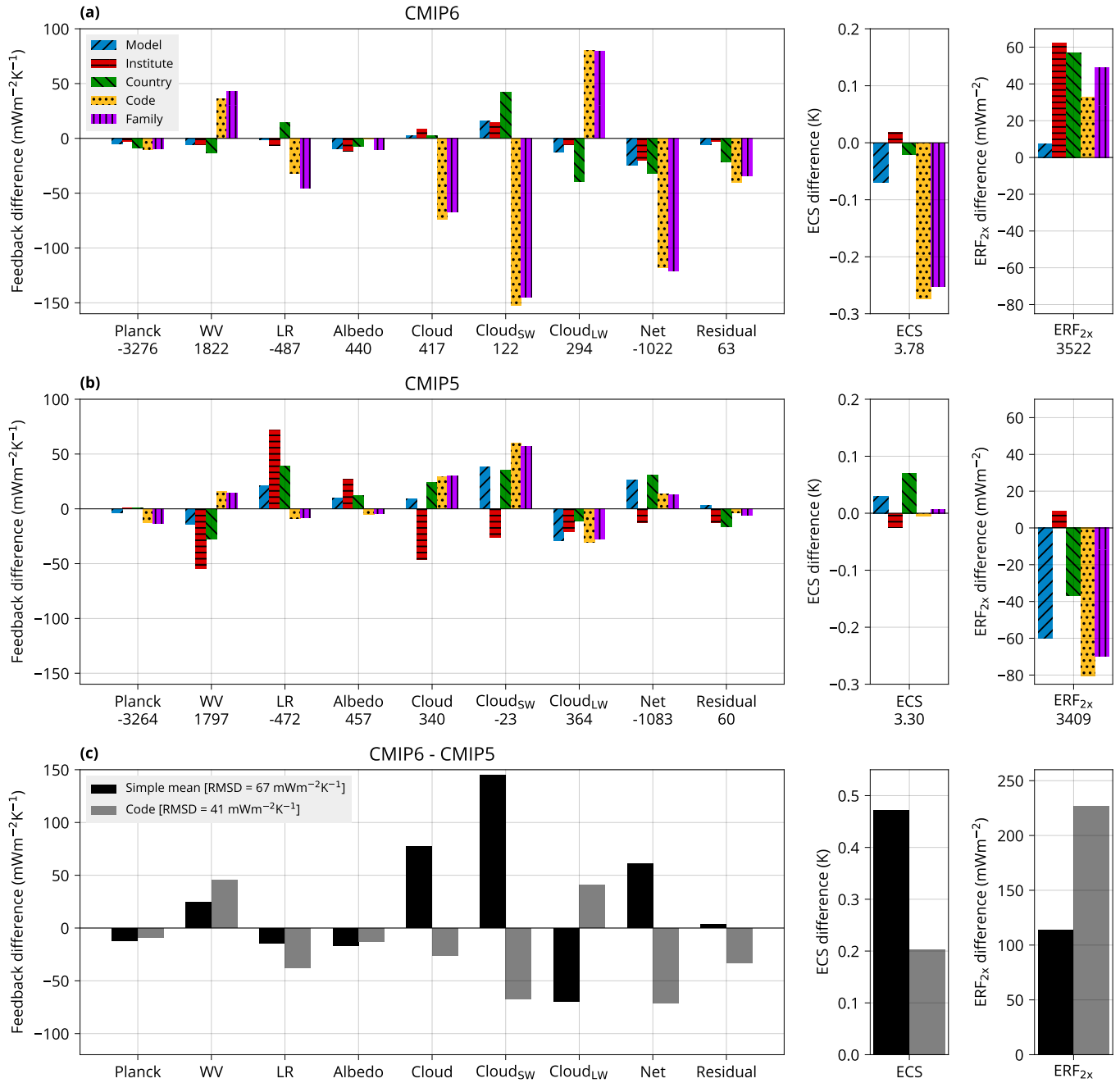


Figure 3. Climate feedbacks, effective climate sensitivity (ECS), and effective radiative forcing (ERF_{2x}) in the Coupled Model Intercomparison Project (CMIP) phases 6 (a) and 5 (b) under different weighting methods (*model*, *institute*, *country*, *ancestry*, and *family*) relative to a *simple mean* (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).

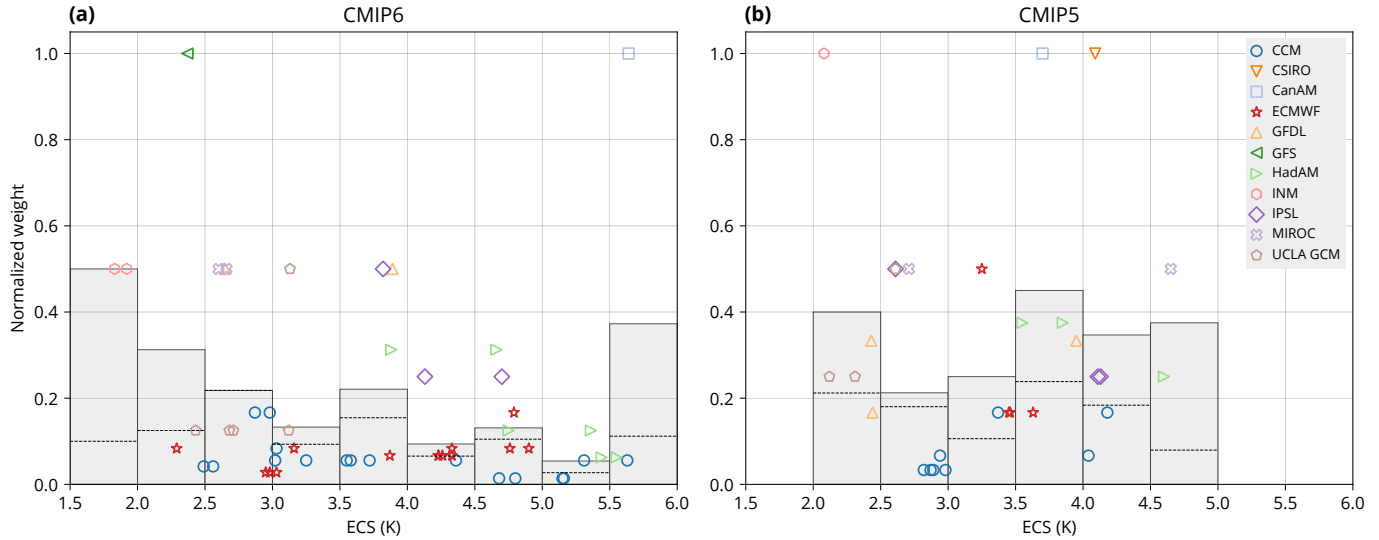


Figure 4. Statistical weights and effective climate sensitivity (ECS) of models in the Coupled Model Intercomparison Project (CMIP) phases 6 (a) and 5 (b) under the *ancestry weighting*. The model weights are normalized so that the maximum value is 1.0. The models are classified by their family, indicated 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.

472 account model code dependence. When the RMSD is calculated for *family weighting* (not
 473 shown in the plot), the RMSD is almost the same as *ancestry weighting* at $42 \text{ mWm}^{-2}\text{K}^{-1}$.
 474 But it is less for the *model weighting* (reduced to $60 \text{ mWm}^{-2}\text{K}^{-1}$), and a slight increase
 475 in RMSD is seen for *institute* (increased to $95 \text{ mWm}^{-2}\text{K}^{-1}$) and *country* (increased to
 476 $79 \text{ mWm}^{-2}\text{K}^{-1}$) weighting. This could mean that only the *ancestry*, *family*, and to a
 477 lesser extent *model weighting* can explain some of the feedback difference between CMIP6
 478 and CMIP5. The result is consistent with the expectation that the *ancestry weighting*
 479 is more suitable than the other types of weighting, which are less strongly related to the
 480 model code genealogy.

481 For ECS and ERF, the differences between weighting methods are also substan-
 482 tial – up to about 0.3 K for ECS and 80 mWm^{-2} for ERF_{2x} in magnitude (Figure 3a,
 483 b). In comparison, the difference in *simple* mean between CMIP6 and CMIP5 is 0.47 K
 484 in ECS and 114 mWm^{-2} in ERF_{2x} , and the standard deviation is 0.73 K and 1.06 K in
 485 ECS (CMIP5 and CMIP6, resp.) and 390 mWm^{-2} and 490 mWm^{-2} in ERF_{2x} (CMIP5
 486 and CMIP6, resp.). The difference in ensemble mean ECS between CMIP6 and CMIP5
 487 becomes much smaller with *ancestry weighting*, falling from 0.47 K (*simple* mean) to 0.20
 488 K (*ancestry weighting*), but the difference in ERF_{2x} is increased from 114 to 226 mWm^{-2} .
 489 Thus, it is possible that a weighting method which accounts for model code dependency
 490 can explain some of the difference in ECS between CMIP5 and CMIP6 as resulting from
 491 an over-representation of models with high ECS in the CMIP6 ensemble.

492 Figure 4 shows model ECS and the statistical weights of models under the *ances-*
 493 *try weighting*. It can be seen that in CMIP6, the model weight is the highest for the low-
 494 est ECS range and progressively lower with increasing ECS (except for the highest ECS
 495 range), due to the fact that models with higher ECS are generally populated by the large
 496 model families HadAM, CCM, and to a lesser extent IPSL and ECMWF, while mod-
 497 els 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*

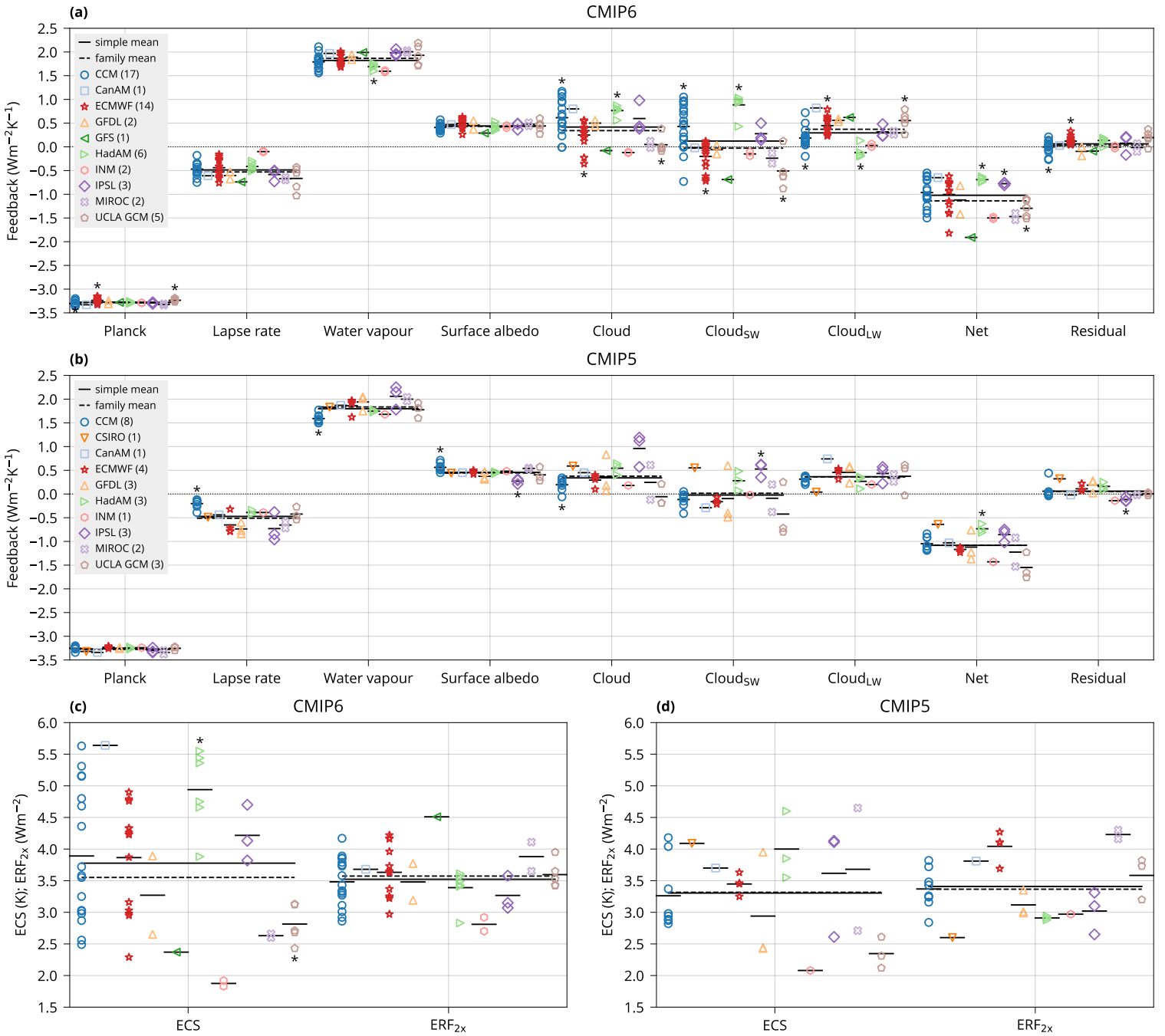


Figure 5. Climate feedbacks, effective climate sensitivity (ECS), and effective radiative forcing (ERF_{2x}) arranged by model family in the Coupled Model Intercomparison Project (CMIP) phases 5 (**b, d**) and 6 (**a, c**). Model family is identified by the oldest ancestor model. In the legend, numbers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

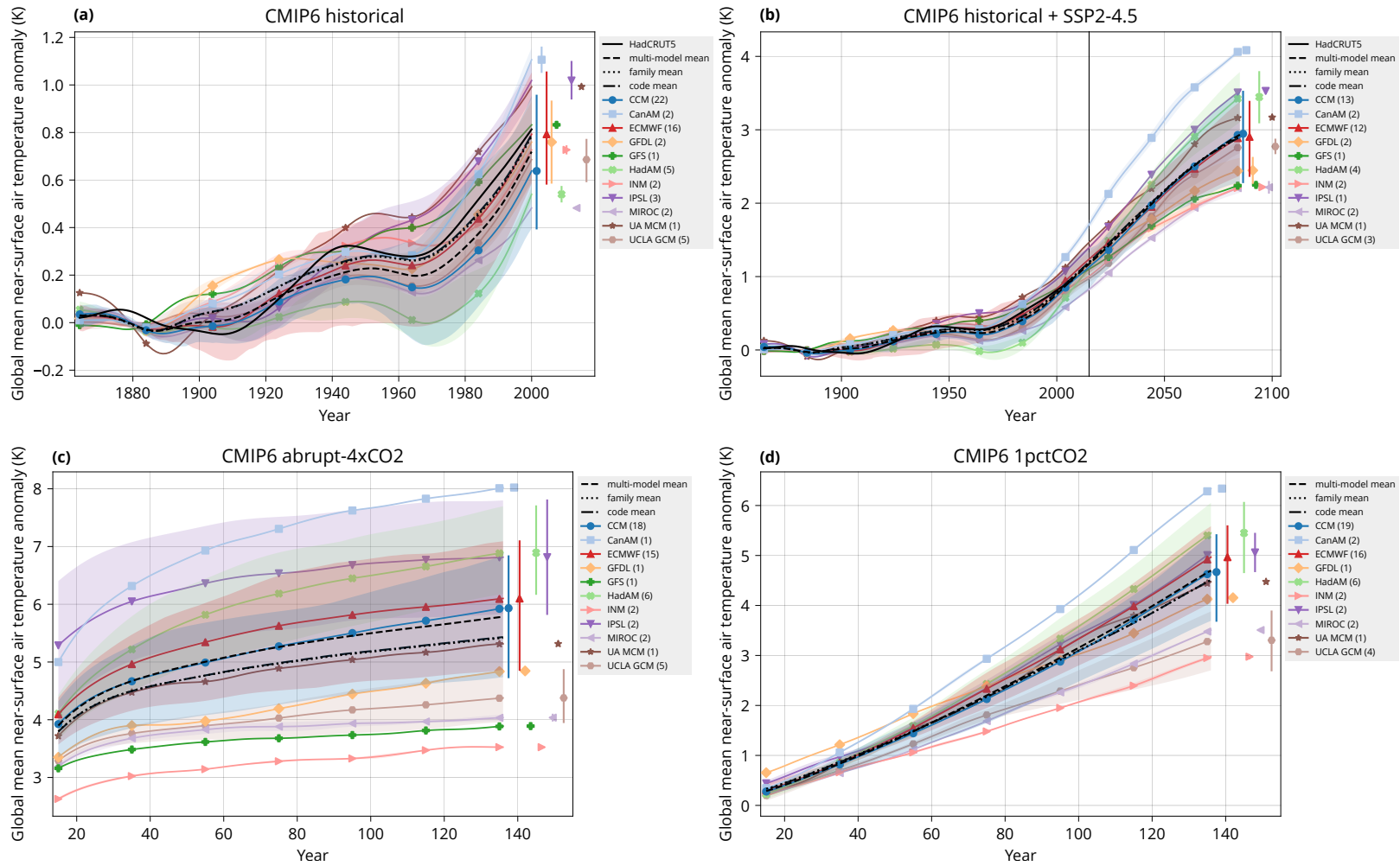


Figure 6. Time series of global mean near-surface temperature in CMIP6 experiments by model family and the *simple* multi-model, *ancestry*, and *family* mean (section 3.2). The model family time series are a *simple* mean of models in the family. The time series are smoothed with a Gaussian kernel with 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 68th 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.

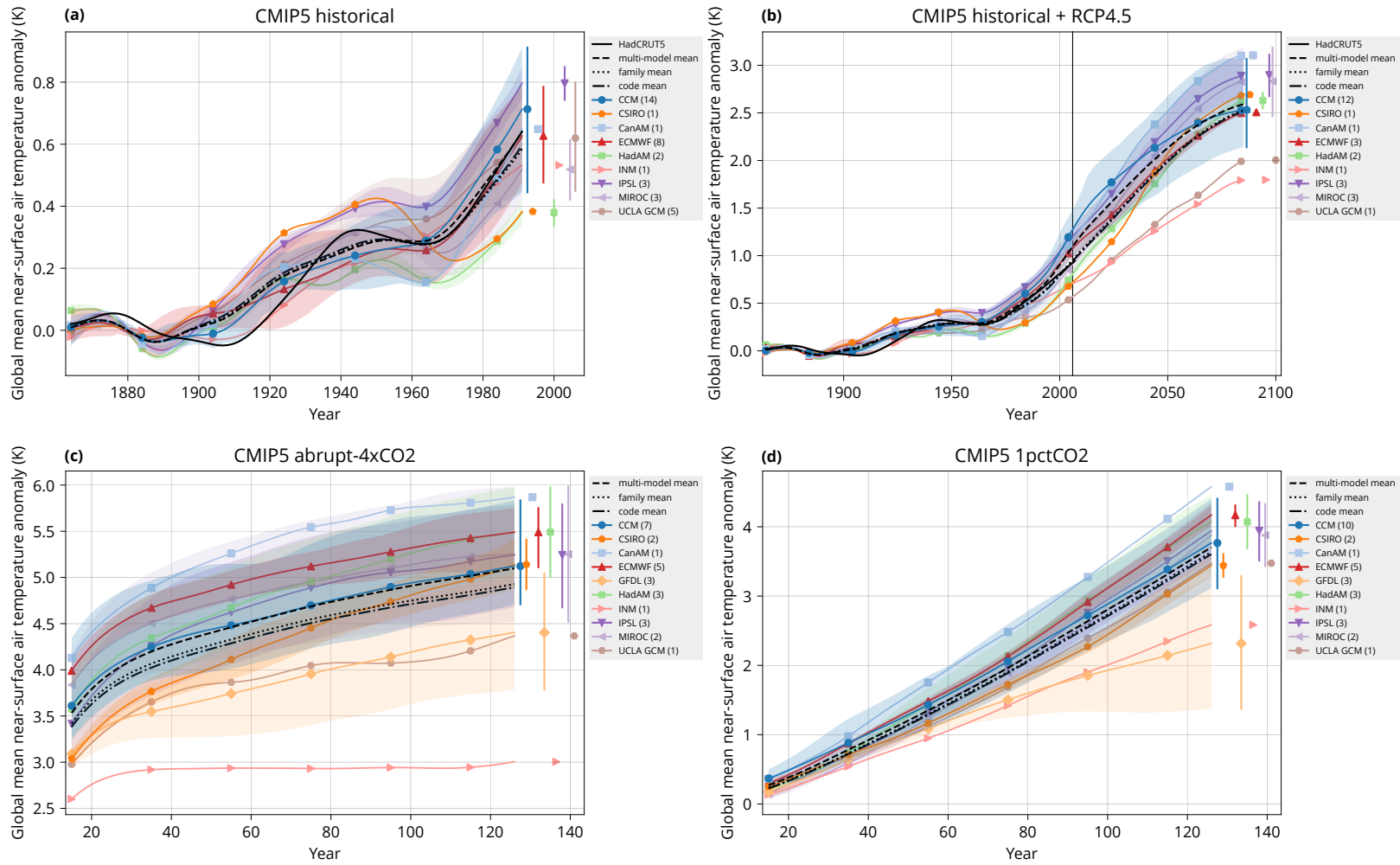


Figure 7. The same as Figure 6 but for CMIP5, and the RCP4.5 experiment instead of SSP2-4.5.

548 time series. Included are all models which provided the necessary data. While some model
 549 families have many members in this analysis, such as CCM (7 to 22 members, depend-
 550 ing on the experiment and CMIP phase), ECMWF (3 to 16 members), HadAM (2 to 6
 551 members), and UCLA GCM (1 to 5 members), other families have less than 4 members,
 552 and therefore it is harder (or impossible) to assess model spread in the smaller families.
 553 The larger families such as CCM and ECMWF exhibit a large spread and a middle-of-
 554 the-range family mean, although the spread of the ECMWF family in the CMIP5 ex-
 555 periments *historical* + RCP4.5 (combined experiments), *abrupt-4xCO2*, and *1pctCO2*
 556 is relatively narrow. The other larger family HadAM has a relatively small spread in most
 557 experiments, consistent with the results of section 4.3. Notably, in the CMIP6 *histor-*
 558 *ical* experiment, HadAM is the coldest of all model families, but becomes the second and
 559 third warmest in the rest of the CMIP6 experiments by the end of the simulation. The
 560 UCLA GCM family of models have consistently relatively low GMST in the CMIP6 *abrupt-*
 561 *4xCO2* and *1pctCO2* experiments, despite the relatively large size of the group (here 4
 562 to 5 members). Model families like MIROC, INM, and CanAM (each containing 2 mem-
 563 bers in the CMIP6 plots, except for CanAM in *abrupt-4xCO2* with only member) have
 564 almost no spread in the CMIP6 experiments, suggesting that the two models in each of
 565 these model families are very similar.

566 The *family* and *ancestry weighted* GMST time series tend to nearly overlap in all
 567 cases, which points to a high degree of outcome similarity between the two types of weight-
 568 ing also noted in the preceding sections. Interestingly, the *family* and *ancestry weighted*
 569 mean is warmer than the *simple* multi-model mean in the CMIP6 *historical* experiment
 570 (in the CMIP5 *historical* experiment it is slightly colder by the end of the simulation)
 571 and also more consistent with observations, whereas in the *1pctCO2* and *abrupt-4xCO2*
 572 experiments it is colder than the *simple* mean (in both CMIP6 and CMIP5). When CMIP6
 573 is compared with CMIP5, model families tend to exhibit similar cold or warm propen-
 574 sity, such as INM, GFDL, UCLA GCM being relatively cold in the non-*historical* exper-
 575 iments, and CanAM, HadAM, IPSL being relatively warm. This suggests that model fam-
 576 ilies tend to maintain their climate sensitivity inclination across model generations.

577 5 Discussion and Conclusions

578 We mapped the code genealogy of 167 models in and related to CMIP3, CMIP5,
 579 and CMIP6 with a focus on the atmospheric component and the atmospheric physics.
 580 We showed that all models can be grouped into 14 model families based on code inher-
 581 itance, although large amounts of code may have been replaced in some models, and there-
 582 fore they are only weakly related to other models in the same family. In addition, we mapped
 583 the institute and country of origin of the models. Some model families, such as CCM,
 584 ECMWF, and HadAM, are particularly large. The CCM-derived models were extensively
 585 forked internationally, most likely due to the open availability of the code. The IFS/ARPEGE
 586 (licensed) code was the basis for many European models. The HadGEM code was shared
 587 internationally within a consortium. Together, these three large model families domi-
 588 nate CMIP6, accounting for 70% of all model runs, an increase from about 50% repre-
 589 sented by the three largest model families in CMIP3 and CMIP5. Based on the code ge-
 590 nealogy, we developed an *ancestry weighting* method, the aim of which was to more fairly
 591 weigh code-related models than a *simple* multi-model mean, thus mitigating structural
 592 model dependence effects in MMEs. We showed that when applied on CMIP5 and CMIP6,
 593 the *ancestry* and *family weighting* produced substantial differences in the climate feed-
 594 backs, sensitivity, and forcing, especially the cloud feedbacks (total, shortwave and long-
 595 wave), ECS, and ERF_{2x} relative to the difference in *simple* mean between CMIP6 and
 596 CMIP5 and relative to the standard deviation of the quantities in CMIP5 and CMIP6.
 597 The *ancestry* and *family weighting* methods produce very similar results. The *ancestry*
 598 and *family weighting* seem to be able to explain some of the difference between CMIP6
 599 and CMIP5 (about 40% RMSD reduction in climate feedbacks, and about 60% RMSD

600 reduction in ECS under the *ancestry weighting*). This suggests that increased contribu-
 601 tions from many code-related models in CMIP6 compared to CMIP5 were able to sub-
 602 stantially affect the *simple* multi-model mean. Applying these methods to analyze cli-
 603 mate feedbacks, sensitivity, and forcing by model family revealed that models in some
 604 families gave narrowly similar results (HadAM and UCLA GCM), and others in some
 605 cases had relatively wide spread but consistently above- or below-mean values (ECMWF
 606 and CSM). This suggests that code similarity in some cases translates to similarities in
 607 climate properties, but in other cases there is a large spread despite model similarity. Lastly,
 608 we analyzed GMST time series in four CMIP6 and CMIP5 experiments, and showed that
 609 models in some larger families (HadAM, and in some cases ECMWF) have similar GMST.
 610 The *family* and *ancestry weighting* showed very similar results – more warming than the
 611 *simple* mean (and closer to observations) in the CMIP6 *historical* experiment and less
 612 warming in the CMIP6 *1pctCO2* and *abrupt-4xCO2* experiments. This suggests that these
 613 methods can partially balance the effect of the over-representation of model families with
 614 multiple similar models, like HadAM. Model families tend to exhibit tendencies toward
 615 greater or lower warming than the MME mean in response to increased CO₂ across the
 616 CMIP generations.

617 A limitation of our method of weighting based on model families or model code ge-
 618 nealogy is that we have not quantified model similarity in other ways than through in-
 619 heritance. We did not make an attempt to quantify model code independence from their
 620 parent models, because there is not enough publicly available information on the source
 621 code. Even if the source code were available, an objective quantification of code inde-
 622 pendence would require a sophisticated new method of code analysis. Some models have
 623 code bases which are more independent from their parent models than others. As a re-
 624 sult, some model families might have members which are almost code-independent from
 625 the rest of the family. For example, it is possible that models which are related in the
 626 genealogy diverged enough from their ancestral models that it would be warranted to
 627 classify them as a separate family. This means that some models can be unjustly under-
 628 weighted because they are grouped together with models to which they do not bear much
 629 resemblance or were developed for a different purpose in mind (discussed below). Over-
 630 coming this limitation would be a relatively difficult task. While it might be possible to
 631 investigate individual schemes and components in models to partially quantify the sta-
 632 tistical distances between related models, it would be difficult to do so objectively. Such
 633 information is also unlikely to be available for all the CMIP participating models. An-
 634 other possibility would be to analyze the code of models to quantify their similarity. A
 635 method of accurately quantifying similarity would necessitate analyzing large code bases,
 636 distinguishing scientific calculations from technical code, accounting for the fact that small
 637 changes in code can produce large differences in model results, and accounting for model
 638 runtime configuration. Emerging methods of code analysis based on deep artificial neural
 639 networks (DANNs) have a potential to be used for this task. DANN-based tools such
 640 as OpenAI Codex (Chen et al., 2021; OpenAI, 2023), GitHub Copilot (GitHub, 2023)
 641 and DeepMind AlphaCode (DeepMind, 2023) have been developed to translate natural
 642 text to computer code. This approach has a potential to be adapted to quantifying code
 643 similarity. However, regardless of the availability of such methods, access to the model
 644 code would be necessary. This is a substantial hurdle given that most model code is closed-
 645 source. Apart from this, the source code of older models (dating back several decades)
 646 might not be readily available even to the current modeling groups, or even preserved
 647 at all. In summary, users of our model code genealogy should be mindful that the pro-
 648 posed weighting methods are only a “first-order” approximation of model similarity, and
 649 they should make an educated choice when selecting models for an analysis or deciding
 650 which models to include in a model family for the purpose of weighting.

651 Structural dependence between code-related models is sometimes reduced by di-
 652 verging purposes of models. We did not make an attempt to quantify this because lim-
 653 itations similar to those mentioned above. The purpose of a model, such as a geograph-

654 ical, process, or quantity focus, is only rarely explicitly stated and it would be difficult
 655 to objectively quantify this divergence. In such case the *family* and *ancestry weighting*
 656 can give too little weight to those models in the same family or branch of the code ge-
 657 nealogy which are substantially different from the rest of the models due to their pur-
 658 pose. One way in which models are divergent within the same family or branch is their
 659 complexity in terms of being an AGCM, AOGCM or ESM (Figure 2). It can be expected
 660 that ESMs are substantially different from a related AOGCM due to the inclusion of the
 661 carbon cycle, vegetation, atmospheric chemistry, biochemistry and other processes. Sim-
 662 ilarly AGCMs, even though rarely participating in CMIP as standalone models, are ex-
 663 pected to differ substantially from related AOGCMs because they do not contain a prog-
 664 nostic ocean component. One way of accounting for this would be to analyze AOGCMs
 665 and ESMs separately. For example, Meehl et al. (2020) note that emissions feedbacks
 666 included in the ESM GFDL-ESM4 (Dunne et al., 2020) reduce ECS compared to its par-
 667 ent AOGCM GFDL-CM4 (Held et al., 2019); GFDL-ESM4 has ECS 3.9 K and GFDL-
 668 CM4 has ECS 2.6 K. In summary, the focus solely on model code inheritance as presented
 669 here does not account for this context, introducing limitations to our weighting meth-
 670 ods.

671 To put our results into a broader perspective, we do not argue against the use of
 672 *simple* multi-model means, or model output and performance weighting methods in gen-
 673 eral, but see the presented weighting methods as complementary to the established meth-
 674 ods. *Simple* means will likely continue to represent a useful default option (as used, for
 675 example, in parts of AR6), but other weighting methods may be increasingly important
 676 due to model duplication in MMEs. It is possible that weighting methods based on model
 677 structure can capture these interdependencies better than methods based on model out-
 678 put. We suggest the family weighting, or a similar technique based on selecting a num-
 679 ber of “independent” model branches from the model code genealogy, as a useful and
 680 easily implemented method of weighting for MME studies, especially if there is an ex-
 681 pectation that model duplication is affecting the results.

682 The presented model code genealogy (Figure 2) can be further extended as more
 683 models become available in future CMIP phases. We provide the Scalable Vector Graph-
 684 ics (SVG) source of this figure so that it can be extended in the future, and all related
 685 code and data are in the supplementary code under an open source license.

686 Our results can facilitate MME assessments, which depend on the knowledge of model
 687 code relations. They provide a complementary approach to the model output dependence
 688 methods presented in previous studies. We have shown that as expected, code-related
 689 models tend to have related climate characteristics, which may help to explain some of
 690 the difference between CMIP5 and CMIP6. Certain model families stand out in terms
 691 of ECS or climate feedbacks, which can help in understanding model differences. This
 692 is especially important given that the model spread in ECS and some climate feedbacks
 693 have increased in CMIP6 relative to CMIP5. A useful method of accounting for depen-
 694 dencies among models is weighting model families equally, which has the benefit of be-
 695 ing simpler to achieve than ancestry weighting. This can be readily employed in MME
 696 assessments if a more fair model weighting is desired.

697 **Appendix A Model Ancestry Weight Calculation**

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

701 Definitions:

- 702 1. *Node* is a single model (AGCM, AOGCM or ESM). It can comprise multiple model
703 runs (configurations or resolutions) submitted to CMIP. Nodes can have one or
704 more parent and child nodes.
- 705 2. *Model run* is a specific model configuration or resolution submitted to CMIP. Some
706 models only have one run in CMIP.
- 707 3. *Group* is a set of nodes with the same model name but different version numbers.
708 In Figure 2, these are connected with horizontal arrows. Group ancestors are all
709 node ancestors of all nodes in the group.
- 710 4. *Root nodes* are nodes which do not have have any ancestors. These are the top-
711 level nodes marked with a thick outline in Figure 2.
- 712 5. *Root groups* are groups which contain a root node.
- 713 6. *Active nodes* and *active model runs* are those which are included in the set of mod-
714 els of interest, i.e. models for which weights are to be calculated.
- 715 7. *Active groups* are groups which contain at least one active node.
- 716 8. *Child node* and *child group* is a direct descendant of its *parent node* or *parent group*.
- 717 9. *Descendant* of a node or group is a direct or indirect (more than one level deep)
718 descendant of the node or group.

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

- 720 1. Groups and nodes which are not active and have no active descendants are removed
721 from the tree.
- 722 2. All nodes and groups are assigned a weight of zero.
- 723 3. All root groups are given the same weight equal to $1/n$, where n is the number
724 of root groups.
- 725 4. For all groups which have already inherited weight from all of their ancestors (or
726 have no ancestors) and are not marked as done, their child groups inherit weight.
727 If the parent group is active, each child group’s weight is incremented by $1/(n+$
728 $1)$, where n is the number of child groups, and the parent group’s weight is set to
729 $1/(n+1)$. If the parent group is not active, each child group’s weight is incremented
730 by $1/n$, and the parent group’s weight is set to zero. The parent group is marked
731 as done.
- 732 5. If all groups are marked as done, continue with Step 6. Otherwise, go back to Step
733 4.
- 734 6. Within each group, active nodes are given weight equal to x/n , where x is the weight
735 of the group and n is the number of active nodes in the group.
- 736 7. For each node, active model runs of the node are given weight equal to x/n , where
737 x is the weight of the node and n is the number of active model runs.

738 Open Research Section

739 Our data processing and visualization code, as well as the associated data are avail-
740 able publicly on GitHub (Kuma, 2022a) and Zenodo (Kuma, 2022b). The version used
741 in our analysis is 1.0.0. The software is licensed under an open source license (MIT), the
742 project internal data files and the output data files are in the public domain (Creative
743 Commons license CC0, <https://creativecommons.org/publicdomain/zero/1.0/>), and
744 the model code genealogy graph images and output plots are licensed under the Creative
745 Commons Attribution 4.0 International license (CC BY 4.0, [https://creativecommons](https://creativecommons.org/licenses/by/4.0/)
746 [.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/)). CMIP5 and CMIP6 model output is publicly available on the
747 Earth System Grid Federation websites (CMIP5, 2022; CMIP6, 2022). The input data
748 for model ECS and climate feedbacks are available publicly from Zelinka (2022). The
749 HadCRUT5 data are available publicly from the Met Office Hadley Centre (2022).

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” (<https://forces-project.eu>) and nextGEMS (<https://nextgems-h2020.eu/>), 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 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., . . . 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](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)*

- 803 [Dataset]. Retrieved from <https://esgf-node.llnl.gov/projects/cmip6/>
 804 (last access: 1 August 2022)
- 805 DeepMind. (2023). *AlphaCode*. Retrieved from <https://alphacode.deepmind.com>
 806 (last access: 27 April 2023)
- 807 Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G.,
 808 ... Zhao, M. (2020). The GFDL Earth System Model version 4.1 (GFDL-
 809 ESM 4.1): Overall coupled model description and simulation characteristics.
 810 *Journal of Advances in Modeling Earth Systems*, 12(11), e2019MS002015.
 811 (e2019MS002015 2019MS002015) doi: 10.1029/2019MS002015
- 812 Edwards, P. N. (2000a). *The agcm family tree*. Retrieved from [http://pne.people
 .si.umich.edu/vastmachine/agcm.html](http://pne.people

 813 .si.umich.edu/vastmachine/agcm.html) (last access: 3 May 2023)
- 814 Edwards, P. N. (2000b). *Atmospheric general circulation modeling: A participatory
 815 history*. Retrieved from [http://pne.people.si.umich.edu/sloan/mainpage
 .html](http://pne.people.si.umich.edu/sloan/mainpage

 816 .html) (last access: 12 August 2022)
- 817 Edwards, P. N. (2000c). Chapter 2: A brief history of atmospheric general cir-
 818 culation modeling. In D. A. Randall (Ed.), *General circulation model de-
 819 velopment* (Vol. 70, pp. 67–90). Academic Press. Retrieved from [https://
 www.sciencedirect.com/science/article/pii/S0074614200800509](https://

 820 www.sciencedirect.com/science/article/pii/S0074614200800509) doi:
 821 10.1016/S0074-6142(00)80050-9
- 822 Edwards, P. N. (2011). History of climate modeling. *WIREs Climate Change*,
 823 2(1), 128–139. Retrieved from [https://wires.onlinelibrary.wiley.com/
 doi/abs/10.1002/wcc.95](https://wires.onlinelibrary.wiley.com/

 824 doi/abs/10.1002/wcc.95) doi: 10.1002/wcc.95
- 825 Edwards, P. N. (2013). Chapter 7: The infinite forecast. In *A vast machine: Com-
 826 puter models, climate data, and the politics of global warming* (pp. 139–186).
 827 The MIT Press.
- 828 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., &
 829 Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project
 830 Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model
 831 Development*, 9(5), 1937–1958. Retrieved from [https://gmd.copernicus
 .org/articles/9/1937/2016/](https://gmd.copernicus

 832 .org/articles/9/1937/2016/) doi: 10.5194/gmd-9-1937-2016
- 833 Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Cald-
 834 well, P., ... Williamson, M. S. (2019, jan). Taking climate model eval-
 835 uation to the next level. *Nature Climate Change*, 9(2), 102–110. Re-
 836 trieved from <https://doi.org/10.1038/s41558-018-0355-y> doi:
 837 10.1038/s41558-018-0355-y
- 838 Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D.,
 839 ... Zhang, H. (2021). The Earth’s energy budget, climate feedbacks, and
 840 climate sensitivity. In *Climate change 2021: The physical science basis.
 841 Contribution of Working Group I to the Sixth Assessment Report of the In-
 842 tergovernmental Panel on Climate Change* (pp. 923–1054). Cambridge Uni-
 843 versity Press, Cambridge, United Kingdom and New York, NY, USA. doi:
 844 10.1017/9781009157896.009
- 845 GitHub. (2023). *Copilot*. Retrieved from <https://github.com/features/copilot>
 846 (last access: 27 April 2023)
- 847 Gjermundsen, A., Nummelin, A., Olivié, D., Bentsen, M., Seland, Ø., & Schulz,
 848 M. (2021, Oct 01). Shutdown of Southern Ocean convection controls long-
 849 term greenhouse gas-induced warming. *Nature Geoscience*, 14(10), 724-
 850 731. Retrieved from <https://doi.org/10.1038/s41561-021-00825-x> doi:
 851 10.1038/s41561-021-00825-x
- 852 Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe,
 853 J. D., ... Zhu, Q. (2019). The DOE E3SM coupled model version 1: Overview
 854 and evaluation at standard resolution. *Journal of Advances in Modeling Earth
 855 Systems*, 11(7), 2089–2129. doi: 10.1029/2018MS001603
- 856 Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S.,
 857 ... Taylor, K. E. (2013). Documenting climate models and their simula-

- 858 tions. *Bulletin of the American Meteorological Society*, *94*(5), 623–627. Re-
 859 trieved from [https://journals.ametsoc.org/view/journals/bams/94/5/](https://journals.ametsoc.org/view/journals/bams/94/5/bams-d-11-00035.1.xml)
 860 [bams-d-11-00035.1.xml](https://journals.ametsoc.org/view/journals/bams/94/5/bams-d-11-00035.1.xml) doi: 10.1175/BAMS-D-11-00035.1
- 861 Houghton, N., Abramowitz, G., Pitman, A., & Phipps, S. J. (2015). Weighting
 862 climate model ensembles for mean and variance estimates. *Climate dynamics*,
 863 *45*(11), 3169–3181. doi: 10.1007/s00382-015-2531-3
- 864 Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J.,
 865 ... Zadeh, N. (2019). Structure and performance of gfdl’s cm4.0 climate
 866 model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691–3727.
 867 Retrieved from [https://agupubs.onlinelibrary.wiley.com/doi/abs/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001829)
 868 [10.1029/2019MS001829](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001829) doi: <https://doi.org/10.1029/2019MS001829>
- 869 Jebeile, J., & Crucifix, M. (2021). Value management and model pluralism in
 870 climate science. *Studies in History and Philosophy of Science*, *88*, 120–127.
 871 Retrieved from [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S003936812100087X)
 872 [S003936812100087X](https://www.sciencedirect.com/science/article/pii/S003936812100087X) doi: 10.1016/j.shpsa.2021.06.004
- 873 Jun, M., Knutti, R., & Nychka, D. W. (2008a). Spatial analysis to quantify nu-
 874 merical model bias and dependence. *Journal of the American Statistical*
 875 *Association*, *103*(483), 934–947. Retrieved from [https://doi.org/10.1198/](https://doi.org/10.1198/016214507000001265)
 876 [016214507000001265](https://doi.org/10.1198/016214507000001265) doi: 10.1198/016214507000001265
- 877 Jun, M., Knutti, R., & Nychka, D. W. (2008b). Local eigenvalue analysis of
 878 CMIP3 climate model errors. *Tellus A: Dynamic Meteorology and Oceanog-*
 879 *raphy*, *60*(5), 992–1000. Retrieved from [https://doi.org/10.1111/](https://doi.org/10.1111/j.1600-0870.2008.00356.x)
 880 [j.1600-0870.2008.00356.x](https://doi.org/10.1111/j.1600-0870.2008.00356.x) doi: 10.1111/j.1600-0870.2008.00356.x
- 881 Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3), 395–404.
 882 doi: 10.1007/s10584-010-9800-2
- 883 Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges
 884 in combining projections from multiple climate models. *Journal of Climate*,
 885 *23*(10), 2739–2758. Retrieved from [https://journals.ametsoc.org/view/](https://journals.ametsoc.org/view/journals/clim/23/10/2009jcli3361.1.xml)
 886 [journals/clim/23/10/2009jcli3361.1.xml](https://journals.ametsoc.org/view/journals/clim/23/10/2009jcli3361.1.xml) doi: 10.1175/2009JCLI3361.1
- 887 Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Genera-
 888 tion CMIP5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194–
 889 1199. Retrieved from [https://agupubs.onlinelibrary.wiley.com/doi/abs/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50256)
 890 [10.1002/grl.50256](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50256) doi: 10.1002/grl.50256
- 891 Krishnan, R., Swapna, P., Choudhury, A. D., Narayansetti, S., Prajeesh, A. G.,
 892 Singh, M., ... Ingle, S. (2021). *The IITM Earth System Model (IITM ESM)*.
 893 arXiv. doi: 10.48550/ARXIV.2101.03410
- 894 Kuma, P. (2022a). *Code accompanying the manuscript "Climate model code*
 895 *genealogy and its relation to climate feedbacks and sensitivity" (Version*
 896 *1.0.0) [Software]*. Retrieved from [https://github.com/peterkuma/](https://github.com/peterkuma/model-code-genealogy-2022/)
 897 [model-code-genealogy-2022/](https://github.com/peterkuma/model-code-genealogy-2022/) (last access: 6 December 2022)
- 898 Kuma, P. (2022b). *Code accompanying the manuscript "Climate model code ge-*
 899 *nealogy and its relation to climate feedbacks and sensitivity" (Version 1.0.0)*
 900 *[Software]*. Zenodo. doi: 10.5281/zenodo.7407118
- 901 Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., & Seland,
 902 Ø. (2022). Machine learning of cloud types in satellite observations and
 903 climate models. *Atmospheric Chemistry and Physics*. (in press) doi:
 904 10.5281/zenodo.7400969
- 905 Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of cli-
 906 mate model pluralism. *Studies in History and Philosophy of Science Part B:*
 907 *Studies in History and Philosophy of Modern Physics*, *41*(3), 253–262. doi: 10
 908 .1016/j.shpsb.2010.07.001
- 909 Lynch, P. (2008). The origins of computer weather prediction and climate
 910 modeling. *Journal of Computational Physics*, *227*(7), 3431–3444. Re-
 911 trieved from [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0021999107000952)
 912 [S0021999107000952](https://www.sciencedirect.com/science/article/pii/S0021999107000952) doi: 10.1016/j.jcp.2007.02.034

- 913 Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research*
 914 *Letters*, 38(8). Retrieved from [https://agupubs.onlinelibrary.wiley.com/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011GL046864)
 915 [doi/abs/10.1029/2011GL046864](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011GL046864) doi: 10.1029/2011GL046864
- 916 Masson-Delmotte, V., et al. (Eds.). (2021). *Climate change 2021: The physical sci-*
 917 *ence basis. Contribution of Working Group I to the Sixth Assessment Report of*
 918 *the Intergovernmental Panel on Climate Change*. Cambridge University Press,
 919 Cambridge, United Kingdom.
- 920 Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B.,
 921 ... Taylor, K. E. (2007). The WCRP CMIP3 multimodel dataset: A new era
 922 in climate change research. *Bulletin of the American Meteorological Society*,
 923 88(9), 1383–1394. Retrieved from [https://journals.ametsoc.org/view/](https://journals.ametsoc.org/view/journals/bams/88/9/bams-88-9-1383.xml)
 924 [journals/bams/88/9/bams-88-9-1383.xml](https://journals.ametsoc.org/view/journals/bams/88/9/bams-88-9-1383.xml) doi: 10.1175/BAMS-88-9-1383
- 925 Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer,
 926 R. J., ... Schlund, M. (2020). Context for interpreting equilibrium cli-
 927 mate sensitivity and transient climate response from the CMIP6 Earth
 928 system models. *Science Advances*, 6(26), eaba1981. Retrieved from
 929 <https://www.science.org/doi/abs/10.1126/sciadv.aba1981> doi:
 930 10.1126/sciadv.aba1981
- 931 Mendlik, T., & Gobiet, A. (2016). Selecting climate simulations for impact stud-
 932 ies based on multivariate patterns of climate change. *Climatic change*, 135(3),
 933 381–393. doi: 10.1007/s10584-015-1582-0
- 934 Met Office Hadley Centre. (2022). *HadCRUT5 [Dataset]*. Retrieved from [https://](https://www.metoffice.gov.uk/hadobs/hadcrut5/)
 935 www.metoffice.gov.uk/hadobs/hadcrut5/ (last access: 12 December 2022)
- 936 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E.
 937 (1953). Equation of state calculations by fast computing machines. *The*
 938 *journal of chemical physics*, 21(6), 1087–1092.
- 939 Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E.,
 940 ... Simpson, I. R. (2021). An updated assessment of near-surface temper-
 941 ature change from 1850: The HadCRUT5 data set. *Journal of Geophysical*
 942 *Research: Atmospheres*, 126(3), e2019JD032361. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361)
 943 agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361
 944 (e2019JD032361 2019JD032361) doi: 10.1029/2019JD032361
- 945 Morrison, M. A. (2021). *The models are alright: A socio-epistemic theory of the*
 946 *landscape of climate model development* (Unpublished doctoral dissertation).
 947 Indiana University, Indiana, United States.
- 948 OpenAI. (2023). *Codex*. Retrieved from <https://openai.com/blog/openai-codex>
 949 (last access: 27 April 2023)
- 950 Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy*
 951 *of Science*, 87(3), 457–477. doi: 10.1086/708691
- 952 Parker, W. S., & Winsberg, E. (2018, Jan 01). Values and evidence: how mod-
 953 els make a difference. *European Journal for Philosophy of Science*, 8(1), 125-
 954 142. Retrieved from <https://doi.org/10.1007/s13194-017-0180-6> doi: 10
 955 .1007/s13194-017-0180-6
- 956 Pennell, C., & Reichler, T. (2011). On the effective number of climate mod-
 957 els. *Journal of Climate*, 24(9), 2358–2367. Retrieved from [https://](https://journals.ametsoc.org/view/journals/clin/24/9/2010jcli3814.1.xml)
 958 journals.ametsoc.org/view/journals/clin/24/9/2010jcli3814.1.xml
 959 doi: 10.1175/2010JCLI3814.1
- 960 Pulkkinen, K., Undorf, S., Bender, F., Wikman-Svahn, P., Doblas-Reyes, F., Flynn,
 961 C., ... Thompson, E. (2022, Jan 01). The value of values in climate sci-
 962 ence. *Nature Climate Change*, 12(1), 4–6. Retrieved from [https://doi.org/](https://doi.org/10.1038/s41558-021-01238-9)
 963 [10.1038/s41558-021-01238-9](https://doi.org/10.1038/s41558-021-01238-9) doi: 10.1038/s41558-021-01238-9
- 964 Pulkkinen, K., Undorf, S., & Bender, F. A.-M. (2022, Nov 18). Values in cli-
 965 mate modelling: testing the practical applicability of the Moral Imagina-
 966 tion ideal. *European Journal for Philosophy of Science*, 12(4), 68. Re-
 967 trieved from <https://doi.org/10.1007/s13194-022-00488-4> doi:

- 10.1007/s13194-022-00488-4
- 968
969 Remmers, J. O., Teuling, A. J., & Melsen, L. A. (2020). Can model structure fami-
970 lies be inferred from model output? *Environmental Modelling & Software*, *133*,
971 104817. Retrieved from [https://www.sciencedirect.com/science/article/
972 pii/S1364815219308436](https://www.sciencedirect.com/science/article/pii/S1364815219308436) doi: 10.1016/j.envsoft.2020.104817
- 973 Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016, apr). Probabilistic program-
974 ming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. Retrieved
975 from <https://doi.org/10.7717/peerj-cs.55> doi: 10.7717/peerj-cs.55
- 976 Sanderson, B. M., Knutti, R., & Caldwell, P. (2015a). Addressing interdependency
977 in a multimodel ensemble by interpolation of model properties. *Journal of Cli-
978 mate*, *28*(13), 5150–5170. Retrieved from [https://journals.ametsoc.org/
979 view/journals/clim/28/13/jcli-d-14-00361.1.xml](https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00361.1.xml) doi: 10.1175/JCLI-D
980 -14-00361.1
- 981 Sanderson, B. M., Knutti, R., & Caldwell, P. (2015b). A representative democ-
982 racy to reduce interdependency in a multimodel ensemble. *Journal of Cli-
983 mate*, *28*(13), 5171–5194. Retrieved from [https://journals.ametsoc.org/
984 view/journals/clim/28/13/jcli-d-14-00362.1.xml](https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00362.1.xml) doi: 10.1175/
985 JCLI-D-14-00362.1
- 986 Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B.,
987 Fisher, R. A., & Knutti, R. (2021). The potential for structural errors
988 in emergent constraints. *Earth System Dynamics*, *12*(3), 899–918. Re-
989 trieved from <https://esd.copernicus.org/articles/12/899/2021/> doi:
990 10.5194/esd-12-899-2021
- 991 Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., & Eyring, V. (2020).
992 Emergent constraints on equilibrium climate sensitivity in CMIP5: do they
993 hold for CMIP6? *Earth System Dynamics*, *11*(4), 1233–1258. Retrieved
994 from <https://esd.copernicus.org/articles/11/1233/2020/> doi:
995 10.5194/esd-11-1233-2020
- 996 Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C.,
997 ... Saha, S. (2017). Practice and philosophy of climate model tuning across
998 six us modeling centers. *Geoscientific Model Development*, *10*(9), 3207–3223.
999 Retrieved from <https://gmd.copernicus.org/articles/10/3207/2017/>
1000 doi: 10.5194/gmd-10-3207-2017
- 1001 Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Har-
1002 greaves, J. C., ... Zelinka, M. D. (2020). An assessment of Earth’s climate
1003 sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4),
1004 e2019RG000678. Retrieved from [https://agupubs.onlinelibrary.wiley
1005 .com/doi/abs/10.1029/2019RG000678](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000678) (e2019RG000678 2019RG000678) doi:
1006 10.1029/2019RG000678
- 1007 Steinschneider, S., McCrary, R., Mearns, L. O., & Brown, C. (2015). The effects of
1008 climate model similarity on probabilistic climate projections and the implica-
1009 tions for local, risk-based adaptation planning. *Geophysical Research Letters*,
1010 *42*(12), 5014–5044. Retrieved from [https://agupubs.onlinelibrary.wiley
1011 .com/doi/abs/10.1002/2015GL064529](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GL064529) doi: 10.1002/2015GL064529
- 1012 Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and
1013 the experiment design. *Bulletin of the American Meteorological Society*, *93*(4),
1014 485–498. Retrieved from [https://journals.ametsoc.org/view/journals/
1015 bams/93/4/bams-d-11-00094.1.xml](https://journals.ametsoc.org/view/journals/bams/93/4/bams-d-11-00094.1.xml) doi: 10.1175/BAMS-D-11-00094.1
- 1016 Touzé-Peiffer, L., Barberousse, A., & Le Treut, H. (2020). The Coupled
1017 Model Intercomparison Project: History, uses, and structural effects on
1018 climate research. *WIREs Climate Change*, *11*(4), e648. Retrieved from
1019 <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcc.648> doi:
1020 10.1002/wcc.648
- 1021 Undorf, S., Pulkkinen, K., Wikman-Svahn, P., & Bender, F. A.-M. (2022, Oct 03).
1022 How do value-judgements enter model-based assessments of climate sensitivity?

- 1023 *Climatic Change*, 174(3), 19. Retrieved from <https://doi.org/10.1007/s10584-022-03435-7> doi: 10.1007/s10584-022-03435-7
- 1024
- 1025 Voosen, P. (2022). ‘Hot’ climate models exaggerate Earth impacts. *Science (New*
- 1026 *York, NY)*, 376(6594), 685–685. doi: 10.1126/science.adc9453
- 1027 Wang, C., Soden, B. J., Yang, W., & Vecchi, G. A. (2021a). Compensation between
- 1028 cloud feedback and aerosol-cloud interaction in CMIP6 models. *Geophys-*
- 1029 *ical Research Letters*, 48(4), e2020GL091024. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091024)
- 1030 agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091024
- 1031 (e2020GL091024 2020GL091024) doi: 10.1029/2020GL091024
- 1032 Williams, J., Morgenstern, O., Varma, V., Behrens, E., Hayek, W., Oliver, H., ...
- 1033 Frame, D. (2016). Development of the New Zealand Earth System Model:
- 1034 NZESM. *Weather and Climate*, 36, 25–44. doi: 10.2307/26779386
- 1035 Winsberg, E. (2012). Values and uncertainties in the predictions of global climate
- 1036 models. *Kennedy Institute of Ethics Journal*, 22(2), 111–137. Retrieved from
- 1037 <https://muse.jhu.edu/pub/1/article/484359> doi: 10.1353/ken.2012.0008
- 1038 Zelinka, M. D. (2022). *GitHub repository mzelinka/cmip56_forcing_feedback_ecs*
- 1039 *[Dataset]*. Retrieved from [https://github.com/mzelinka/cmip56_forcing](https://github.com/mzelinka/cmip56_forcing_feedback_ecs)
- 1040 [_feedback_ecs](https://github.com/mzelinka/cmip56_forcing_feedback_ecs) (last access: 3 August 2022)
- 1041 Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M.,
- 1042 Ceppi, P., ... Taylor, K. E. (2020). Causes of higher climate sensitivity
- 1043 in CMIP6 models. *Geophysical Research Letters*, 47(1), e2019GL085782.
- 1044 Retrieved from [https://agupubs.onlinelibrary.wiley.com/doi/abs/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782)
- 1045 [10.1029/2019GL085782](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782) (e2019GL085782 10.1029/2019GL085782) doi:
- 1046 [10.1029/2019GL085782](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782)