Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity

Julien Boe

AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2017GL076829

Key Points:
- The relationship between the number of components shared by GCMs and the proximity of their results is quantified precisely
- No component clearly appears as more important in that context
- Different strategies to deal with interdependency in multimodel ensembles a priori are discussed

Supporting Information:
- Supporting Information S1
- Table S1

Correspondence to:
J. Boé,
boe@cerfacs.fr

Citation:
Boé, J. (2018). Interdependency in multimodel climate projections: Component replication and result similarity. *Geophysical Research Letters*, *45*, 2771–2779. https://doi.org/10.1002/2017GL076829

Received 15 DEC 2017
Accepted 15 FEB 2018
Accepted article online 7 MAR 2018
Published online 23 MAR 2018

Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity

Julien Boé[1]

[1]CECI, Université de Toulouse, CNRS, Cerfacs, Toulouse, France

Abstract Multimodel ensembles are the main way to deal with model uncertainties in climate projections. However, the interdependencies between models that often share entire components make it difficult to combine their results in a satisfactory way. In this study, how the replication of components (atmosphere, ocean, land, and sea ice) between climate models impacts the proximity of their results is quantified precisely, in terms of climatological means and future changes. A clear relationship exists between the number of components shared by climate models and the proximity of their results. Even the impact of a single shared component is generally visible. These conclusions are true at both the global and regional scales. Given available data, it cannot be robustly concluded that some components are more important than others. Those results provide ways to estimate model interdependencies a priori rather than a posteriori based on their results, in order to define independence weights.

1. Introduction

Global climate models (GCMs), as all models, are not perfect. The multimodel approach has been the main strategy followed to estimate how these imperfections affect climate projections, with the development of large intercomparison projects, in particular the Coupled Model Intercomparison Projects (CMIP, e.g., Eyring et al., 2016; Meehl et al., 2007; Taylor et al., 2012).

At the core of the multimodel approach lies the basic idea that if the results of an additional model B are close to the ones of a model A, then our confidence in the results of A is reinforced. Obviously, it is only true insofar as A and B are not nearly identical. The more similar B and A are, the less informative are the results of B regarding the realism of those of A. Note that defining precisely what "near identical" or "similar" exactly means in that context is a difficult question, as discussed later.

The notion of model independence broadly encapsulates these ideas but has very seldom been precisely defined. Annan and Hargreaves (2017), who propose a statistical definition of independence, are a notable exception. A condition of independence would be that for all the subjective choices that are not strongly constrained by the fundamental principles of physics (and, maybe, to a lesser extent by observations regarding physical parameterizations) the modeling groups make their choices independently. As GCMs often share a common history (Edwards, 2011) and scientific literature leads to a diffusion of ideas, it is not the case. Some even argue, as Mazzocchi and Pasini (2017), that as all the climate models share the same fundamental equations and principles—they all are dynamical models—they cannot be considered as independent in any case.

For CMIP5 (Taylor et al., 2012), about 22 modeling groups have done coordinated climate projections with around 40 GCMs . These models sometimes share entire components (Table 9.A.1 in Flato et al., 2013, or Table S1 in the supporting information) with therefore a strong level of interdependency. Despite that, the "model democracy" (i.e., "one model, one vote"; Knutti, 2010), which implicitly assumes their independence, has prevailed until today. Even the last Intergovernmental Panel on Climate Change report generally uses projections from all the available GCMs in the analyses shown (see the number of models in many figures of Collins et al., 2013).

The lack of independence of CMIP models and the limits of models democracy have been pointed a long time ago (e.g., Knutti, 2010) and are now universally accepted. The fact that model democracy is still widely used today despite that is likely the result of the lack of universal and totally satisfactory alternatives.

A growing number of studies have recently tackled the question of how to deal with interdependency in multimodel projections. The approaches proposed can be divided into two main families. In the a posteriori approaches, the proximity of GCMs results or of their errors is used to quantify a posteriori their interdependencies (e.g., Abramowitz & Gupta, 2008; Bishop & Abramowitz, 2013). Independence weights can then be computed based on the proximity of GCMs results as in Sanderson et al. (2015, 2017) and Steinschneider et al. (2015) and used to derive ensemble statistics. In the a priori approaches, the independence of models is judged a priori, based only on the knowledge of their codes and not of their results.

The two approaches are not equivalent. Strongly dependent models are expected to produce very similar results (e.g., Masson & Knutti, 2011; Knutti et al., 2013), but as argued by Annan and Hargreaves (2017), models with similar results are not necessarily not independent. In fact, supposing that the similarity of GCMs results necessarily implies a lack of independence is somewhat contradictory with the premises of the multimodel approach. If when a model B gives almost the same results as a model A we automatically decide that A and B are not independent, the similarity of the results of B with those of A will never reinforce our confidence in A, which is the cornerstone of the multimodel approach as noted earlier.

The a priori approach, advocated by Annan and Hargreaves (2017), therefore, may appear more satisfactory from a theoretical point of view. It has received little practical use yet, likely because of the difficulty to define independence weights a priori based on the similarity of the codes.

A few studies have proposed simple approaches in the "a priori" framework, for example, the "institutional democracy" or "one climate modeling group, one vote" proposed by Leduc et al. (2016). As GCMs from the same institution often show major code similarities, the authors propose to give the same final weight to each institution rather than to each GCM. Annan and Hargreaves (2017) propose a general a priori method to compute independence weights based on whether or not the models come from the same institution.

As noted by the authors of the previous studies, deciding whether or not two GCMs are independent based on their institutions is just a first step. A better knowledge of how code similarity impacts GCMs results is needed to go forward. This is the objective of this study, following previous works on the links between structural similarities and the proximity of results (e.g., Annan & Hargreaves, 2017; Leduc et al., 2016; Masson & Knutti, 2011; Pennell & Reichler, 2011; Steinschneider et al., 2015) but working at the finer level of components.

2. Data and Methods

Forty GCMs from the CMIP5 project (Taylor et al., 2012) are studied. Each GCM is characterized by its four main components: atmosphere, ocean, land surface, and sea ice models (Table S1). The number of different submodels for the four components (Table S1) is much smaller than the number of GCMs, pointing to an important level of component replication.

Temperature, precipitation, and sea level pressure (SLP) from multiple members of the historical and radiative concentration pathways 8.5 (RCP8.5) experiments are analyzed. Climatological averages on the 1970–1999 period and future changes between 2070–2099 and 1970–1999 are studied.

Pairwise correlations or root-mean-square errors (RMSEs) between GCMs or their errors are often used to study GCMs interdependencies (e.g., Jun et al., 2008; Knutti et al., 2010; Pennell & Reichler, 2011). Here pairwise spatial RMSEs between all pairs of simulations are computed. The pairwise RMSEs are then flagged into different categories, according to the shared components of the corresponding pairs of GCMs. The pairwise RMSEs between GCMs that share 0, 1, 2, 3, or 4 components are respectively grouped in the C0, C1, C2, C3, and C4 categories. The C4 category corresponds to GCMs that share their four main components but differ by either their resolution, secondary components (e.g., interactive atmospheric chemistry or biogeochemistry; "secondary" is not used here to qualify a priori the impact of the corresponding structural changes on model results but because, historically, these components were generally included later in GCMs), or possibly tuning parameters.

Some GCMs may indeed share a component but use different values for some parameters, based on different tuning strategies (Hourdin et al., 2016). Here even with different tuning parameters, the shared components are considered "identical." While it is not totally satisfactory because different tuning parameters may clearly impact model results (e.g., Mauritsen et al., 2012), the lack of documentation of tuning strategies, as noted by Hourdin et al. (2016), makes it almost impossible to follow another approach.

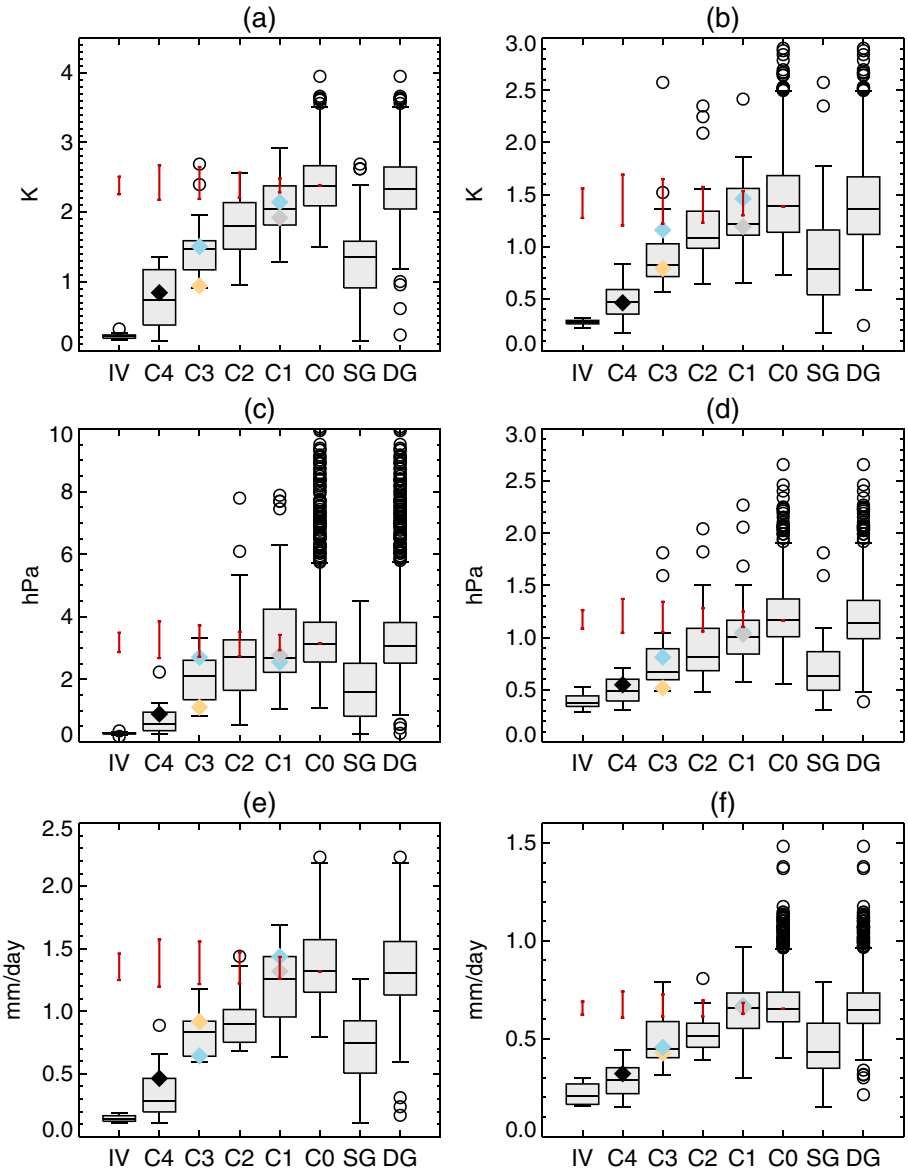

Figure 1. Distribution of pairwise root-mean-square errors (RMSEs) for different categories (section 2). The box-and-whisker plots show the median, interquartile range, and either the minimum and maximum, or the 75th centile plus 1.5 interquartile range and the 25th centile minus 1.5 interquartile range if there are some values outside this range. In that case, these outlier values are shown with circles. The distribution of median pairwise RMSEs obtained with the random sample test described in section 2 are shown in red. It corresponds to the null hypothesis: no difference with C0. The colored diamonds show the median pairwise RMSE for different subcategories. For C4, the black diamonds correspond to models that only differ by their resolution. For C3, the yellow (blue) diamond corresponds to models that do not share their oceanic (atmospheric) component. For C1, the blue (gray) diamond corresponds to models that only share the ocean (ice) component. (a) Climatological temperature (K) on the 1970–1999 period, (b) future change in temperature (K) between 2070–2099 and 1970–1999. (c and d) Same as (a) and (b) for sea level pressure (hPa). (e and f) Same as (a) and (b) for precipitation (mm/day).

The category IV (internal variability) corresponds to the pairwise RMSEs between members from the exact same GCM, that is, between simulations that only differ by initial conditions. Two other categories indicate whether the GCMs come from the same modeling group (SG) or from different modeling groups (DG). The sample size of each categories is given in Table S2.

All the possible pairwise RMSEs between the n members of a model A and the m members of a model B are computed. The ensemble mean of the $n \cdot m$ pairwise RMSEs is then computed to reduce the impact of internal

variability. As the number of members is not always large, internal variability still impacts pairwise RMSEs within all categories.

Deciding whether two GCMs share the "same" component or not is a difficult question. Some GCMs indeed use different versions of a same component (Table S1), and deciding when structural changes justify to consider two versions as different is partly subjective (see also the extended discussion in the supporting information (SI)).

As a general rule, if the first version number is different (e.g., CAM4 and CAM5), the components are considered different. If the second version number is different (e.g., CICE4 and CICE4.1) the components are considered identical. Some case-by-case exceptions based on the actual differences of codes between the versions are introduced and discussed in the SI.

These choices are likely not perfect and partly subjective. Someone with a better knowledge of the differences of codes could have made different choices. Note that it is crucial not to analyze the results of the GCMs (or to use the knowledge gained from a previous analysis) to decide whether or not they share the same component, in order to avoid circular reasoning and to follow a true a priori approach. As some subjectivity necessarily remains when deciding whether two components are similar, sensitivity tests with tighter and looser rules for the definition of similarity are described in the SI.

To test whether the differences of RMSEs between a given category and C0 (no shared component category) are significant, a simple nonparametric test based on random resampling is used. For each category (C1, C2, etc.) the sample size of pairwise RMSEs is known (Table S2). The same number of pairwise RMSEs is then randomly chosen within the C0 category, and the median is computed. This operation is repeated 10,000 times. The distributions of the median RMSE obtained thanks to this procedure are shown with red error bars (5–95% range) in Figure 1.

3. Results

A clear relationship exists between the number of components shared by two GCMs and the proximity of their results, for climatological averages and future changes of all the variables studied (Figure 1). The median pairwise RMSE for pairs of models that share no component (C0) is between 2 or 5 times greater than the median pairwise RMSE for pairs of GCMs that share their four main components and differ by the resolution or secondary components (C4). GCMs with three shared components (C3) are also much closer together than models with no shared component (median RMSE between 1.5 or 2 times smaller than that of C0). Differences of RMSEs between C2 and C1 are generally clear. The impact of a single shared component (C1), although generally small, is visible and significant in most cases.

As the sample sizes are not always large (Table S2), some results may be model dependent. For example, large pairwise RMSEs for climatological SLP lead to a large third quartile for C1 and many outliers for C0 (Figure 1c). A closer inspection reveals that these large RMSEs are generally due to the three Institut Pierre-Simon-Laplace (IPSL) models. Climatological SLPs over the Himalaya, Greenland, and Antarctica in these models are outliers compared to the rest of the CMIP5 models (not shown.)

The impact of internal variability is weak for present-day averages (i.e., much smaller than the impact of a change in resolution or addition of secondary components characterized in C4). It is generally greater for future changes, notably for SLP and precipitation changes. In these cases, the C4 median pairwise RMSEs are only 1/3 or 1/4 greater than the ones of IV.

As expected given the strong component replication within same-group models (Table S1), the pairwise RMSEs of SG are generally much smaller than that of DG (Figure 1). The median pairwise RMSE for SG is generally close to that of C3, consistent with the fact that on average SG models share 3.20 components.

The DG pairwise RMSEs distributions are generally close to that of C0 in terms of interquartile ranges and medians, because the number of shared components by GCMs from different groups is still moderate. Eighty-nine percent of DG pairwise RMSEs are indeed also in C0. However, regarding the lower tails of RMSE distributions, clear differences between DG and C0 are seen (Figure 1). Small pairwise RMSEs often seen in DG (outliers or end of the inferior whisker) are not seen in C0. It is particularly visible for climatological temperature, SLP, precipitation, and precipitation changes (Figure 1). The small pairwise RSMEs in DG but not in C0

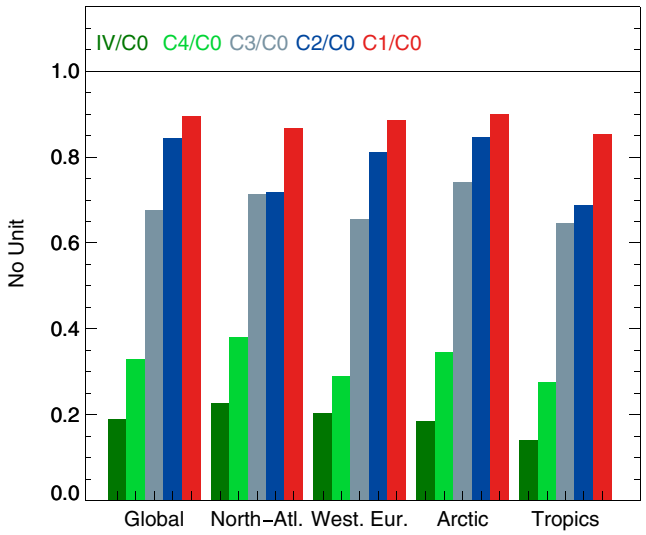

Figure 2. Ratio of the mean pairwise root-mean-square errors (RMSEs) of future temperature change for each category given by the colored bars on the mean pairwise RMSE of C0. The RMSEs are calculated on different regions of the world given on the *x* axis. The boundaries of the North Atlantic domain are 30°N–70°N, −65°E–0°E. The boundaries of the Western Europe domain are 37°N–65°N, −10°E–20°E. Only land points are considered. The Tropics are defined as the zone between −20°N and 20°N. The Arctic is defined as the zone with latitudes greater than 70°N. IV in dark green, C4 in light green, C3 in dark gray, C2 in purple, and C1 in red.

are due to intergroup component replication. This result therefore highlights the interest to go further than the institutional democracy.

To assess whether the strong impact of component replication is still discernible at regional scales, the same analysis as in Figure 1b is done for different regions. The ratio between the mean pairwise RMSEs of each category and the mean pairwise RMSEs of C0 are shown (Figure 2).

For global temperature change, the mean pairwise RMSEs for IV (C4, C3, C2, and C1) is respectively equal to 20% (33%, 68%, 85%, and 90%) of the mean pairwise RMSE of C0. The results for the different regions are very similar to global results. The regional ratios of RMSEs indeed do not deviate by more than ±15% from the ratios obtained for global temperature change (Figure 2). The impact of component replication on the similarity of GCMs results is therefore also strong at the regional scale.

The previous analyses show that a relationship exists between the number of components shared by two GCMs and the proximity of their results. The nature of the component is not accounted for by these analyses, even if all components might not be equally important.

The interest of assessing whether all the components are equally important in shaping the similarity of GCMs response is twofold. First, it is necessary to better judge a priori the independence of two GCMs based on the similarity of their components. Second, it may help to better understand which processes are the most important in shaping the intermodel spread in a given context.

To assess the role of individual components, the C3 (only one component is different) and C1 categories (only one component is identical) are studied. The pairwise RMSEs between two GCMs have also been tagged according to their only shared component for C1 or their only different component for C3 and the ensemble medians computed (colored diamonds in Figure 1). The results are only shown when the sample size is greater than 5 (Table S2). In most cases, the sample sizes are small and the results of this analysis should be considered with caution.

No component clearly appears as systematically more important, as the associated differences of pairwise RMSEs are generally small, given the small sample sizes. For climatological SLP for example, the pairwise RMSEs of C3 are slightly greater when the atmosphere is different than when the ocean is different. Opposite results are obtained for climatological precipitation (Figures 1c and 1e).

As the role of individual components could differ regionally, a regional analysis for temperature change is performed (Figure 3). First, not surprisingly, important spatial variations of the mean pairwise RMSEs for C3 (Figure 3a) and for C1 (Figure 3b) are seen. The pairwise RMSEs are generally much higher in the Arctic, where the intermodel spread is greater.

Except in the North Atlantic, the pairwise RMSEs for GCMs that only differ by one component (Figure 3a) are slightly larger when the atmosphere is different than when the ocean is different, especially in the Tropics and over land, suggesting a somewhat more important role of atmospheric models in driving future temperature change there.

When only one component is identical (C1), pairwise RMSEs are slightly larger when the GCMs share the same ocean component than when they share the same sea ice component, especially in the Arctic, pointing logically to an important role of the ice model there (Figure 3b).

The sample sizes for the analyses on the respective roles of the different components are small, sometimes very small, and few models are involved. Results are therefore very likely to be model dependent, and the relative importance of the different components suggested by the previous analyses should not be overinterpreted. Actually, the most surprising result is perhaps the general absence of strong hierarchy between the different components. For example, the role of the ice component in the Tropics does not not seem less

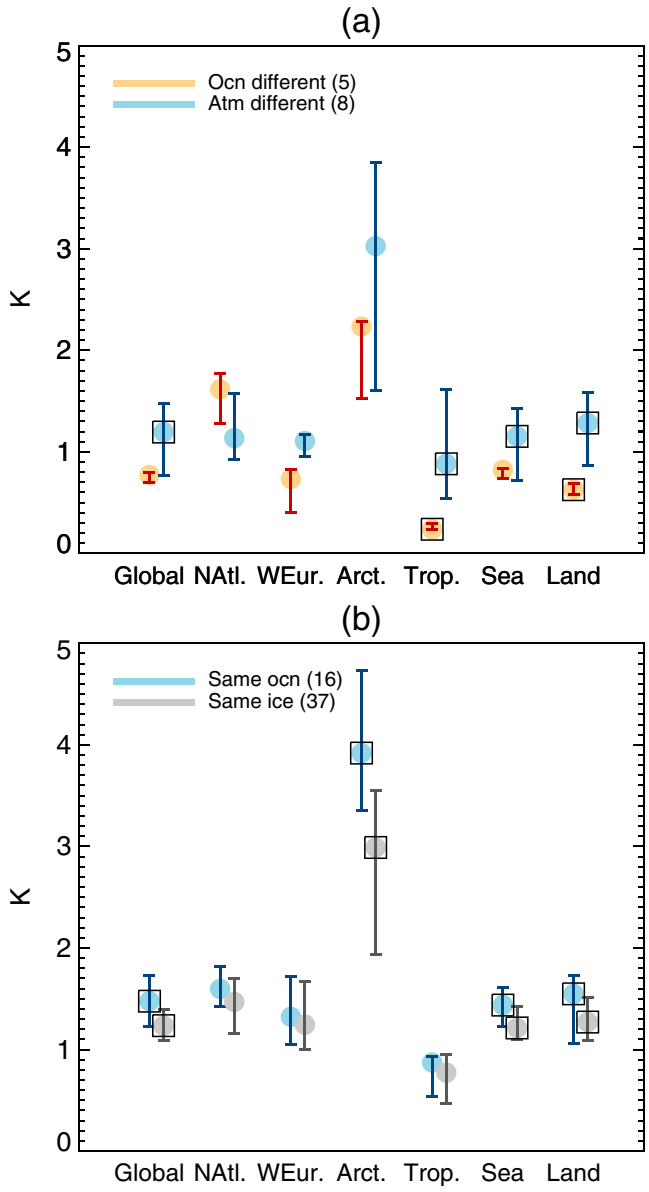

Figure 3. Pairwise root-mean-square errors (RMSEs) for future temperature change over different regions for (a) some subsamples of the C3 category and (b) some subsamples of the C1 category (see section 2). In (a), GCMs that only differ by the atmosphere (ocean) component are shown in blue (yellow). In (b), GCMs that only have the ice (ocean) component in common are shown in gray (blue). The mean is given by the circle, and the whiskers show the 25th and 75th centiles. The black square corresponds to the result of a test similar to the one described at the end of section 2 except that the random selection is done within the C3 category for (a) and within C1 for (b). A black square is drawn when the pairwise RMSE is outside the 5–95% range of the mean pairwise RMSE obtained with 10,000 random selections. Sea and land areas correspond respectively to global averages for sea-only and land-only points. The other regions are defined as in Figure 2.

important than the one of ocean. This is a reminder of the holistic nature of the climate system and the great importance of interregion and intercomponent interactions.

4. Discussion

The previous results raise questions regarding the best direction for the development of multimodel ensembles to go further than ensembles of opportunity (Tebaldi & Knutti, 2007). Developing new GCMs with components already used in other models seems to have been a tendency from CMIP3 to CMIP5. It is likely not an optimal choice from a pure multimodel projection perspective. Based on this paper's results, one could argue that it would be better to focus on the development of fewer GCMs, but with new components, rather than assembling new GCMs with already existing components, adding secondary components, or using different resolutions. Note, however, that GCMs are not only used in the pragmatic goal to provide climate service-oriented projections but also to better understand the climate system, its variability, etc. Component replication may be useful in this context.

This study is only focused on independence weights. Quality weights (e.g., Boé & Terray, 2015; Sanderson et al., 2015) are not dealt with as they are largely different questions. They are not totally unrelated yet. Bishop and Abramowitz (2013), for example, show that both independence and quality weights naturally emerge when weighting multiple models to minimize the distance to observations. Additionally, a component used in many GCMs likely benefits from more human resources for its development and evaluation. One could therefore expect its results to become increasingly more realistic version after version, resulting in an incentive for other groups to use it, in a kind of "Darwinian selection" of components. It is interesting to note that Sanderson et al. (2015), who use both quality and a posteriori independence weights, found an intermodel anticorrelation between them. It could, however, be a result of the definition of independence weights a posteriori, based on the proximity of GCMs results. Indeed, imagine a totally unrealistic GCM because of an important numerical bug: with an a posteriori approach, this GCM would receive a very large independence weight because its results would be far from all the other GCMs. It would also receive a very small quality weight, as it would also be far from the observations, hence the potential existence of a link between quality and a posteriori independence weights. In an a priori approach, this GCM would not necessarily receive a particularly large independence weight, as its results do not matter in this framework. No anticorrelation between independence and quality weights is therefore expected in the a priori approach, except if the idea of Darwinian selection of components has some truth.

5. Conclusion

As discussed in this paper, it is more satisfactory from a theoretical point of view to assess model interdependency a priori based on the similarity of their codes, rather than a posteriori based on the similarity of their results. The practical implementation of the a priori framework is, however, complex.

In this study, it is shown that structural similarities generally clearly increase the proximity of GCMs results. Qualitatively, this result is anything but surprising (e.g., Annan & Hargreaves, 2017; Masson & Knutti, 2011;

Pennell & Reichler, 2011; Steinschneider et al., 2015), but the interest of the present study is to provide a precise quantification of this phenomenon, at the finer level of the component.

The impact of a single shared component on pairwise RMSEs is visible for almost all the variables studied here, and the more components are shared, the stronger the reduction of RMSE is. The differences caused by the simple change of resolution or addition of secondary components are generally small, not necessarily much bigger than the impact of internal variability for future changes. These results are true for different regions of the world.

Some details of these results are sensitive to the exact definition of similarity, as shown by the sensitivity tests in the SI. However, the main conclusions of this work remain robust.

There is at present generally no strong reason to consider one component as more important than another, as far as the four main components—ocean, atmosphere, land, and ice—are concerned. A larger ensemble and/or a higher level of component replication would be necessary to reach stronger (and potentially different) conclusions.

The results presented in this study can be used in practice to deal with interdependency issues in multimodel projections. A basic approach would be to totally forbid component replication, as generally even the replication of a single component has a visible impact. One could also decide to accept a certain level of component replication, based on the reduction of RMSEs that one judges acceptable. Based on our results, the presence of GCMs that share one component, maybe two, in the pruned ensemble might be considered acceptable.

These simple approaches are not optimal, as even if GCMs with replicated components are not independent, they are not totally identical. A better approach might be the "component democracy," whereby each different component would be given the same overall weight in the ensemble. The weight of a GCM would be the combination of the weights corresponding to each of its components. Another possibility would be to extend the approach proposed by Annan and Hargreaves (2017) to derive independence weights but at the replicated component level rather than the group level.

Even if an approach based on intermodel component replication is an interesting step forward, it is still crude and has some limits. First, the choice of considering two components identical or not remains partly subjective and the choices made in this study mainly based on version numbers are not necessarily the best. Additionally, the impact of tuning is not considered. Some components may be considered identical in this work but use different parameters, which may be a source of important differences. A better documentation of tuning in GCMs would be necessary to go further.

Second, different (with the definition used in this study) components often share identical parameterization schemes and are therefore themselves not independent. The models that share no component are therefore not really independent. A way forward would be to expand this paper's framework to work at the level of parameterization, that is, to flag pairwise RMSEs according to the replication of parameterization schemes. This approach would necessitate efficient ways to extract the necessary information, with a complete and standardized documentation of all GCMs. The Common Metadata for Climate Modelling Digital Repositories (METAFOR) project (Moine et al., 2014) and now Earth System Documentation (https://es-doc.org/) are important steps in this direction. Ultimately, one could also imagine using some tools for the automatic comparison of codes (at the algorithmic level) to build an objective measure of model independence.

Acknowledgments
This work has been supported by the French National Research Agency (ANR) in the framework of its JCJC program (ECHO, decision ANR 2011 JS56 014 01). I acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and I thank the climate modeling groups that developed the models listed in supporting information Table S1 of this paper for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Abramowitz, G., & Gupta, H. (2008). Toward a model space and model independence metric. *Geophysical Research Letters*, *35*, L05705. https://doi.org/10.1029/2007GL032834

Annan, J. D., & Hargreaves, J. C. (2017). On the meaning of independence in climate science. *Earth System Dynamics*, *8*, 211–224. https://doi.org/10.5194/esd-8-211-2017

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevag, A., Seland, O., et al. (2013). The Norwegian Earth System Model, NorESM1-M—Part 1: Description and basic evaluation of the physical climate. *Geoscientific Model Development*, *6*, 687–720. https://doi.org/10.5194/gmd-6-687-2013

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

Boé, J., & Terray, L. (2015). Can metric-based approaches really improve model climate projections? The case of summer temperature change in France. *Climate Dynamics*, *45*(7-8), 1913–1928.

Chen, H., Zhou, T., Neale, R. B., Wu, X., & Zhang, G. J. (2010). Performance of the new NCAR CAM3.5 in East Asian summer monsoon simulations: Sensitivity to modifications of the convection scheme. *Journal of Climate*, *23*, 3657–3675. https://doi.org/10.1175/2010JCLI3022.1

Chylek, P., Li, J., Dubey, M. K., Wang, M., & Lesins, G. (2011). Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM2. *Atmospheric Chemisty and Physics Discussion*, *11*, 22,893–22,907. https://doi.org/10.5194/acpd-11-22893-2011

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., et al. (2011). Development and evaluation of an Earth-System model—HadGEM2. *Geoscientific Model Development*, *4*, 1051–1075.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., et al. (2013). Long-term climate change: Projections, commitments and irreversibility. In T. F. Stocker, et al. (Eds.), *Climate change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York: Cambridge University Press.

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., et al. (2011). The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. *Journal of Climate*, *24*, 3484–3519. https://doi.org/10.1175/2011JCLI3955.1

Dufresne, J.-L. ., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., et al. (2013). Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5. *Climate Dynamics*, *40*(9-10), 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Edwards, P. N. (2011). History of climate modeling. *Wiley Interdisciplinary Reviews: Climate Change*, *2*, 128–139. https://doi.org/10.1002/wcc.95

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

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In T. F. Stocker, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York: Cambridge University Press.

Gent, P. R., Yeager, S. G., Neale, R. B., Levis, S., & Bailey, D. A. (2010). Improvements in a half degree atmosphere/land version of the CCSM. *Climate Dynamics*, *34*, 819–833. https://doi.org/10.1007/s00382-009-0614-8

Gordon, H. B., O'Farrell, S. P., Collier, M. A., Dix, M. R., Rotstayn, L. D., Kowalczyk, E. A., et al. (2010). The CSIRO Mk3.5 Climate Model (Technical Report No. 21, 62 pp.). Aspendale, Vic. Australia: The Centre for Australian Weather and Climate Research.

Hazeleger, W., Wang, X., Severijns, C., Ştefaănescu, S., Bintanja, R., Sterl, A., et al. (2012). EC-Earth V2.2: Description and validation of a new seamless earth system prediction model. *Climate Dynamics*, *39*, 2611–2629. https://doi.org/10.1007/s00382-011-1228-5

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2016). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*, 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1

Hunke, E. C., & Lipscomb, W. H. (2010). CICE: The Los Alamos sea ice model, documentation and software user's manual, version 4.1 (Technical Report LACC-06-012). Los Alamos, NM: T-3 Fluid Dynamics group, Los Alamos National Laboratory. Retrieved from http://climate.lanl.gov/models/cice

Jun, M., Knutti, R., & Nychka, D. W. (2008). Spatial analysis to quantify numerical model bias and dependence. *Journal of the American Statistical Association*, *103*, 934–947.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, *102*(3-4), 395–404.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, *23*, 2739–2758. https://doi.org/10.1175/2009JCLI3361.1

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

Leduc, M., Laprise, R., de Elia, R., & Separovic, L. (2016). Is institutional democracy a good proxy for model independence? *Journal of Climate*, *29*, 8301–8316.

Li, L., Lin, P., Yu, Y., Wang, B., Zhou, T., Liu, L., et al. (2013). The flexible global ocean-atmosphere-land system model, grid-point version 2: FGOALS-g2. *Advances in Atmospheric Sciences*, *30*, 543–560. https://doi.org/10.1007/s00376-012-2140-6

Marsh, D. R., Mills, M. J., Kinnison, D. E., Lamarque, J., Calvo, N., & Polvani, L. M. (2013). Climate change from 1850 to 2005 simulated in CESM1(WACCM). *Journal of Climate*, *26*, 7372–7391. https://doi.org/10.1175/JCLI-D-12-00558.1

Mazzocchi, F., & Pasini, A. (2017). Climate model pluralism beyond dynamical ensembles. *Wiley Interdisciplinary Reviews: Climate Change*, *8*, 477. https://doi.org/10.1002/wcc.477

Martin, G. M., Bellouin, N., Collins, W. J., Culverwell, I. D., Halloran, P. R., Hardiman, S. C., et al. (2011). The HadGEM2 family of Met Office Unified Model climate configurations. *Geoscientific Model Development*, *4*, 723–757.

Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*, L08703. https://doi.org/10.1029/2011GL046864

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*, M00A01. https://doi.org/10.1029/2012MS000154

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., et al. (2007). The WCRP CMIP3 Multimodel Dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*, 1383–1394.

Merryfield, W. J., Lee, W., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., et al. (2013). The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Monthly Weather Review*, *141*, 2910–2945. https://doi.org/10.1175/MWR-D-12-00216.1

Moine, M.-P., Valcke, S., Lawrence, B. N., Pascoe, C., Ford, R. W., Alias, A., et al. (2014). Development and exploitation of a controlled vocabulary in support of climate modelling. *Geoscientific Model Development*, *7*, 479–493. https://doi.org/10.5194/gmd-7-479-2014

Pennell, C., & Reichler, T. (2011). On the effective number of climate models. *Journal of Climate*, *24*, 2358–2367. https://doi.org/10.1175/2010JCLI3814.1

Neale, R. B., Chen, C.-C., Gettelman, A., Lauritzen, P. H., Park, S., Williamson, D. L. et al. (2012). Description of the NCAR Community Atmosphere Model (CAM 5.0) (Tech. Rep. NCAR/TN-486+STR). Boulder, Colo: National Center for Atmospheric Research. Retrieved from http://www.cesm.ucar.edu/models/cesm1.0/cam/docs/description/cam5_desc_save.pdf

Qiao, F., Song, A., Bao, Y., Song, Y., Shu, Q., Huang, C., et al. (2013). Development and evaluation of an Earth System Model with surface gravity waves. *Journal of Geophysical Research: Oceans*, *118*, 4514–4524. https://doi.org/10.1002/jgrc.20327

Rotstayn, L. D., Jeffrey, S. J., Collier, M. A., Dravitzki, S. M., Hirst, A. C., Syktus, J. I., et al. (2012). Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: A study using single-forcing climate simulations. *Atmospheric Chemistry and Physics*, *12*, 6377–6404. https://doi.org/10.5194/acp-12-6377-2012

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

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, *10*, 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., et al. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*, 141–184. https://doi.org/10.1002/2013MS000265

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

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498.

Verronen, P. T., Andersson, M. E., Marsh, D. R., Kovács, T., & Plane, J. M. C. (2016). WACCM-D — Whole Atmosphere Community Climate Model with *D*-region ion chemistry. *Journal of Advances in Modeling Earth Systems*, *8*, 954–975. https://doi.org/10.1002/2015MS000592

Voldoire, A., Sanchez-Gomez, E., Salas y Mélia, D., Decharme, B., Cassou, C., Sénési, S., et al. (2013). The CNRM-CM5.1 global climate model: Description and basic evaluation. *Climate Dynamics*, *40*, 2091–2121. https://doi.org/10.1007/s00382-011-1259-y

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., et al. (2010). Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *Journal of Climate*, *23*, 6312–6335. https://doi.org/10.1175/2010JCLI3679.1

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., et al. (2011). MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, *4*, 845–872. https://doi.org/10.5194/gmd-4-845-2011

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., et al. (2012). A new global climate model of the Meteorological Research Institute: MRI-CGCM3 Model description and basic performance. *Journal of the Meteorological Society of Japan. Ser. II*, *90A*, 23–64. https://doi.org/10.2151/jmsj.2012-A02