# The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations

By **K. Alexander and S. M. Easterbrook**

## Response to Reviewers

This document is colour-coded as follows:

- Comments by reviewers are in **blue**.
- Our responses are in **black**.
- Blocks of text we have added to the manuscript are in **red**.

## Anonymous Referee #1

Received and published: 9 February 2015

"The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations" investigates six CMIP5 model configurations and two EMICAR5 configurations. The paper is based on ideas presented by the authors at the AGU conference in 2011. The authors introduce a set of diagrams to visualise the dependencies between model components in the selected Earth system models. All contributing model versions are analysed following the same principles which allows for an immediate intercomparison of software architecture. The additional information about the relative size in terms of source code of the different model model components makes these 8 diagrams a felicitous representation of the general software design. The idea is worth to be published and the diagrams have the potential to become a standard ingredient for every model description paper. Some statements made by the authors require some more proof (for details see below). Some paragraphs would benefit from being rewritten as the reasoning and relevance for this paper is not obvious. Conclusions are too vague.

Specific comments

(1.1) P354, l 16ff: Many individual modelling groups have tried to find simple graphical representations of their own code (HadGEM2-ES see Fig 3 in Collins et al., 2011, MPI-ESMLR see Fig. 1 in Giorgetta et al., 2012, for CESM1 see Fig 2 in Hurrell et al., 2013). Are these in your terminology examples of Bretherton diagrams? It needs to be worked out more clearly what modelling groups have done so far and why the approach taken by the authors is a better one (standardised view, provides relevant information rather than dressed up pictures). Already from the three examples provided above one may guess that some figures are more self-explanatory than others in terms of architectural design.

This is a good point. As the Referee points out, there do exist examples of graphical (block-diagram) views of individual models in the literature. We have clarified this, and expanded our discussion of existing architectural diagrams in section 2. The relevant paragraph now reads:

While intercomparisons of skill scores and climatological patterns are important, these results suggest we need more insight into the nature of similarities and differences between models. The above approaches compare the outputs of the models, but tend to treat the models themselves as black boxes. There are very few representations of the high level designs of global climate models. The Bretherton diagram is perhaps the best known visualization (see Figure 3 of NASA Advisory Council (1986)), although it represents an idealized schematic of Earth system processes, rather than the specific design of any model (Cook, 2013). Other architectural diagrams have been created for individual models (see Figure 3 of Collins et al. (2011), Figure 1 of Giorgetta et al. (2013), and Figure 1 of Hurrell et al. (2013)). However, such diagrams are *ad hoc* conceptual views, and don't claim to be comprehensive, nor to accurately reflect the model codes. They therefore make it hard to compare model architectures in a systematic way. A standardized visualization of architectural structure applied to many models would be more useful for intercomparison, and may even help explain observed similarities between model outputs. For example, if the models share significant subcomponents, this would affect the diversity in a multi-model ensemble.

(1.2) P 354,l 20: some reference could be added to explain what a Bretherton diagram is: Figure 2b in Earth System Science Overview: A Program for Global Change, NASA Advisory Council. Earth System Sciences Committee, 1986?

We have included this citation in the paragraph quoted in (1.1).

(1.3) P 354, l 25: I claim that models can be architecturally similar but be still divers in terms of geoscience aspects. HadGEM2-ES, IPSL-CM5A-LR, and MPI-ESM-LR are very similar in architecture but still quite divers in terms of science as they do not share many (if any) components. See also my comment for P364, l2 ff.

While it's certainly true that models with similar architecture might still be very different scientifically, our argument goes beyond just the arrangement of components and coupling. By "architecture", we mean not just the topography of the model's structure, but also the relative sizes of components, and whether components are re-used from other models.

Note also that the architecture we show in the diagrams is only two levels deep for ease of viewing (components + selected subcomponents). In reality, there are many more levels of architecture that we don't illustrate. The diagrams we show in this paper only give a first-order visualisation of architecture, meaning that some models may seem more architecturally similar than they really are. Finding useful ways of visualizing additional layers is challenging, but may be suitable for future work.

However, this point requires deeper analysis than would make sense for the current study, so we have edited the paragraph quoted in (1.1) to avoid any claim that architecturally similar models would necessarily be scientifically similar. We have also modified our discussion of the relationship between architecture and model climatology in the discussion section to clarify that the relationship is a hypothesis – see our response to point (1.13) below.

(1.4) P 354, l 27 ff: While the authors put forward geoscientific arguments in the previous paragraphs, here they make a short detour into the development of coupler software. While there is nothing wrong about the idea of sharing and reusing model infrastructure (not only for coupling) I do not see the relevance of this statement for this paper. Using coupler A versus coupler B does not have a direct impact on the quality of science that we get out of an Earth system model. I understand Randall (2011) the way that he is addressing the coupling algorithms but not coupling software.

Our paper is not just about "the quality of science that we get out of an Earth system model". We are also interested in the qualities of software architecture that enable efficient development of the science code.

We argue that coupling software is far more important than most scientists generally assume, and that one of the advantages of our diagrams is that they make visible the infrastructure code that is otherwise invisible in scientific descriptions of the models. Because the coupling software represents a significant portion of the code base, it is also a significant part of the development effort of a climate model. Yet this is not code that can easily be handed over to software engineers to build. The specialists who build and maintain this infrastructure code at most modeling centres need expertise in the scientific domains of the software components, in the engineering choices that affect performance on HPC platforms, and, most importantly for this paper, knowledge of how the coupling architecture can facilitate or hamper the ability to address particular scientific questions with the coupled model.

The coupler also indirectly affects the scientific output of models, as we argue in a new paragraph added to section 2:

Coupling software also has an indirect effect on the scientific output of Earth system models, particularly by influencing development pathways. The design of a coupler largely determines the difficulty of using a new scientific component: whether it can be directly linked to the coupler (and if so, how easily?) or if it has to be nested within the atmosphere or ocean components. The coupler design strongly influences parallelization, particularly load balancing between components, and ultimately what simulations can be run within the given constraints on computing power. This will influence the kinds of scientific questions modellers decide to pursue, which in turn further impacts model development. In coupled model experiments, the scientific phenomena of interest (e.g. Earth system feedbacks) tend to cross the boundaries of individual model components. Hence the coupler design can have an unexpected influence on the fidelity of the physical processes in the model. In addition, the coupler design can affect the ease with which scientists can explore how well the model captures large-scale processes (such as ENSO) that cross model component boundaries.

(1.5) P 355, l 9 ff: As already stated in the introduction I appreciate the way how the authors address the general architecture of a given Earth system model, and I have no objection against the statement made in the last sentence of this para, the demand for a comparative analysis. However, I cannot follow the logics that lead to this statement. The text raises the expectation that the authors will now present a top-down analysis, but then the authors only present a top analysis and do not go down deep into the code as it would be necessary for an investigation, comparison and revision e.g. of a coupling algorithm (not coupling software).

We have deleted “top-down” in this sentence, and we have added the following text to section 2 to explain how our study represents a step towards the larger goal:

In response to this observation, we argue that a comparative analysis of the architecture of Earth system models is necessary. The analysis we present in this paper is a first step towards this goal - we focus on the design decisions represented by the top few levels of the dependency tree for each model, without extending the analysis to the low-level routines. Our method of identifying modules based on dependency structure could be applied recursively, leading to much larger and more complex diagrams. We have not pursued this further in current study due to the immense amount of code involved. Our study therefore represents a first step towards a top-down analysis of model architecture at all levels.

(1.6) P 355, l 23: Selection criteria are missing. Why do the authors select those six out of 45 model configurations, why three models from a single country (USA) but none from Asia or Australia?

We have added the following paragraph to section 3:

We could only analyse models where we had access to both the complete source code, *and* a contact at the given institution who was willing to help us pre-process the code and answer questions about the model. We generally relied on our existing contacts, which meant that the resulting models were not geographically representative of the CMIP5 and EMICAR5 participants. However, the variety of component structures and complexity levels found in these eight models suggests that we have sampled across a wide range of CMIP5 and EMICAR5 model architectures.

(1.7) P 359, l 26: If components share the same grid the nesting of those components is probably also more efficient performance wise (as MPI messages are saved) and it eliminates the problem of load balancing between those nested components or enhances the problem of load balancing as there is less freedom in distributing processes.

We have added a sentence to this paragraph acknowledging the implications of component nesting for parallelization:

When two components share the same grid (spatial discretization), nesting them in this manner is much less complicated than routing them through the coupler. It also leads to a simpler, albeit less flexible, parallelization scheme. This approach retains the historical paradigm of Atmosphere-Ocean GCMs (AOGCMs) rather than comprehensive Earth System Models (ESMs), even if the model contains all the processes found in an ESM.

(1.8) P 361, l 15ff: Line count may correlate with code complexity. But I can have numerous physical or dynamical processes implemented in my model code like different advection schemes or radiation code while only one scheme is selected via Fortran-if at run time. The software stack can become incredibly complex (even after the preprocessing step) while the really active model code can still remain remarkably simple. Can the authors comment on this?

The models in our study tend to strip out these unused options during pre-processing, so they are not present during compilation or at run-time. Most models in our study use CPP (C Pre-Processor), which encloses blocks of code in #ifdef statements rather than Fortran if statements; the contents of these #ifdef blocks are either selected or removed prior to compilation based on the options that have been selected.

Only two models in our study (CESM and MPI) may still contain unused code in the manner suggested. CESM uses a preprocessing system other than CPP, where saving the fully extracted code is not possible. The developers of this model trimmed down the code as much as possible before sending it to us. MPI-ESM requires further preprocessing for each experiment (eg CMIP5 historical runs vs RCP simulations) beyond the preprocessing required to extract the configuration MPI-ESM-LR. The developers of this model sent us the code for MPI-ESM-LR including options for all experiments, but they assured us that the amount of duplicated/unused code was minimal. Therefore, the line counts for CESM and MPI may be slightly inflated, but we do not think these issues have significantly biased our results. To acknowledge this point, we have added the following footnote to section 3:

CESM and MPI could not be fully preprocessed for our analysis. Their line counts in Figures 1, 7, and 9 may be slightly inflated; however, we do not think this has significantly biased our results.

(1.9) Is the finding of Herraiz et al. (2007) applicable to Earth system model code?

Herraiz et al demonstrated that different measures of software complexity all correlate well with line count. We have not directly tested this for Earth system model code, although it might be worth doing so as part of a broader investigation of code metrics for scientific code.

We do know that the finding is robust across a range of open source projects, and Earth system model codes share many of the important attributes of such open source projects. Most notably, two of the Earth System models for which we have obtained historical data exhibit the same steady long-term linear growth in size of the code base that characterizes open source projects, over periods as long as fifteen years of model development. Part of this analyses is included in Easterbrook and Johns 2009, as we have now discussed in section 4.3 (the rest is, as yet, unpublished):

Over the development history of a climate model, the line count tends to grow linearly. For example, Easterbrook and Johns (2009) showed that the UK Met Office model grew steadily from 100,000 lines of code in 1993 to nearly 1 million by 2008. The bulk of this code growth is due to addition of new geophysical processes to the model, and an increase in sophistication of the processes that are already incorporated.

The simplest explanation of a long-term linear growth trend in line count is that it represents a steady accumulation of new science in the model, as would be predicted by Herraiz et al's findings. The main challenge to this explanation is that we found one significant discontinuity in the growth of the code base of Had-GEM, when the old ocean component was replaced with NEMO, and it's not clear whether NEMO is a "simpler" model. This example suggests that one has to be careful in using line count as a proxy for complexity when comparing models build by different labs, rather than for analyzing the distribution of complexity within a single model.

(1.10) P362, l 1 ff: What is the knowledge we gain from this paragraph. The more source code a component has the more complex it is? What is the benefit of including poorly understood processes? Does a more complex model deliver a better climate? Is there any insight this analysis can provide about scientific quality of model components, model systems and model results?

The question of whether a more complex model delivers a better climate simulation is hotly contested across the modeling community, and we certainly don't intend to answer it here. We are also not making any claims about the relationship between complexity and scientific quality, and the Referee's point about inclusion of poorly understood processes underscores this.

The point we are making is that the distribution of complexity within a coupled model system does reveal interesting patterns about where a particular lab has invested time and effort, and hence in which parts of the earth system they are likely to have pools of expertise in, and what sort of scientific questions they address. For example, a model with a highly developed active carbon cycle model (such as IPSL-CM5A) should be well-suited to long-term analysis of paleoclimate, while a model with a detailed atmospheric chemistry component (such as HadGEM2-ES) should be better suited to short-term studies of air quality and weather systems. Such observations could be cross-checked with studies of scientific fidelity, although quantitative skill metrics for specific uses of a coupled model do not yet exist.

(1.11) P 363, l 20 ff: The logics of this paragraph is not clear to me. It starts with "organized collaborations between institutions", moves to the OASIS coupler which is built in Toulouse (at CERFACS, to my knowledge with only financial support by CNRS). Is this really a good example for a multi-institutional effort? MPI-ESM-LR is given as an example which uses OASIS. HadGEM(2-ES) is highlighted as a model which “consists almost entirely of in-house components (except for the UKCA atmospheric chemistry)”.How does this compare to the MPI-ESM-LR whose components are almost entirely developed in-house but for the coupler? The message (if any) of this paragraph is more confusing than helpful.

OASIS and NEMO both originated as single-institution components, but have transitioned to multi-institution components, through the support of EU funding and multi-lateral agreements between labs to support the development efforts. The HadGEM example is intended to illustrate that labs which have preferred to develop in-house in the past are now moving more towards the collaborative model. We have edited this paragraph to clarify:

In recent years, there have also been organized collaborations between institutions to build shared components with high levels of scientific complexity. These components are then included in several coupled modelling systems, and typically can also be run in stand-alone configurations. For example, the ocean model NEMO (Madec, 2008; Vancoppenolle et al., 2008), which was originally developed at IPSL, is now incorporated into several European models, and its ongoing development is now managed by a consortium of five European institutions. IPSL (Figure 6) and MPI (Figure 7) both use the OASIS coupler (Valcke, 2013), which was originally developed by scientists from the French institutions CERFACS and CNRS, and is now supported by a collaborative EU-funded research project. Other models are also moving in this direction. For example, the version of HadGEM (Figure 5) included in this study consists almost entirely of in-house components (the UKCA atmospheric chemistry subcomponent is the only major piece of code developed externally), but has now incorporated OASIS, NEMO, and CICE into its next release (Hewitt et al., 2011).

(1.12) P 364, l 16: The factor of 20 one probably gets when comparing EMICs with a full ESM? Is this a fair comparison? A simple box model is likely to be even smaller. But what is the message here, that adding more processes to the model system increases the number of lines of code? Hm.

We have updated the text to include this measurement restricted to the six GCMs (factor of 7), and to clarify that the factor of 20 requires comparing a GCM to an EMIC. We have chosen to retain the factor of 20 statement because we believe it is an interesting and useful illustration of how different GCMs and EMICs are in terms of total complexity, and should help a non-expert reader understand the distinction between these classes of model:

Finally, climate models vary widely in complexity, with the total line count varying by a factor of 20 between the largest GCM and the smallest EMIC we analyse (Figure 9). Even when restricting this comparison to the six GCMs, there is still a factor of 7 variation in total line count.

(1.13) P 364, l 20: The authors state that similarities in architecture lead to a similarity in the simulated climate. I cannot deduce such from Masson and Kutti (2011). On the contrary, IPSL-CM5A-LR and MPI-ESM are similar in architecture but differ in the simulated climate. Can the authors give some evidence for their believe and hypothesis?

This is a similar point to that made in (1.3) above, and our response is the same. It is certainly only a hypothesis, but one that our study offers at least weak support for. We have edited this paragraph to include quotes from Knutti’s 2013 paper that help to explain the intent of this discussion:

Our analysis also offers new insights into the question of model diversity, which is important when creating multi-model ensembles. Masson and Knutti (2011) and Knutti et al. (2013) showed that models from the same lab tend to have similar climatology, even over multiple model generations. We believe this can be explained, at least in part, in terms of their architectural structure and the distribution of complexity within the model. As Knutti et al. (2013) suggest, “We propose that one reason some models are so similar is because they share common code. Another explanation for the similarity of successive models in one institution may be that different centers care about different aspects of the climate and use different data sets and metrics to judge model ‘quality’ during development." Our analysis offers preliminary evidence to support both of these hypotheses. We hypothesize further that the relative size of each component within an Earth system model indicates the relative size of the pool of expertise available to that lab in each Earth system domain (once adjustments are made for components imported from other labs). The availability of different areas of expertise at each lab may provide a sufficient explanation for the clustering effects reported by Masson and Knutti (2011) and Knutti et al. (2013).

(1.14) P 365, l 2 ff: Even though the diagrams do help very much to get an overview of the general design (which is useful and helpful) I am not convinced that the diagrams help to understand the inner workings of climate models. Statements made by the authors are perfectly right. Then the authors speculate about potential usage of their approach. At conferences the authors have communicated their idea about how to visualise those internal structures of Earth system models for some years now. Thus, I would have been glad to read about concrete examples where their approach has already brought some light into the dark mystery of source code and how this has already helped climate scientists to understand each others source code.

We have a growing body of informal evidence that our architectural diagrams are valuable to the scientific community. Each presentation at conferences and workshops has generated a high level of interest in the audience, particularly among scientists who work directly with the models we analyse. They appreciate seeing the diagram representing their institution’s model, but are often much more interested seeing in the diagrams for other models. Seeing the top-level architectural designs of models they have never worked with allows them to see how other institutions have taken different approaches to address the same challenges. Sometimes these comparisons lead to surprises: in particular, GFDL’s strategy of routing all atmosphere-ocean fluxes through the sea ice component is so unknown to scientists outside GFDL that several have initially insisted that our diagram was mistaken.

At least three scientists have used our diagrams in their own presentations, and there may be others who we have not heard about (earlier versions of our diagrams were available freely online). We have also received several reports of eager discussion within research groups that have come across our diagrams. This work has also been useful for public communication: most notably, Scott K. Johnson included one of our diagrams in an article explaining how climate models are built, published on the popular-science website Ars Technica: http://arstechnica.com/science/2013/09/why-trust-climate-models-its-a-matter-of-simple-science/

Based on the reactions we have witnessed, we are confident in the value of our architectural diagrams as scientific communication tools. However, we are reluctant to include this evidence in the current study, as it is necessarily informal. A more rigorous approach would require formal interviews, ethics approval, and potentially the writing of a second manuscript suited to a different journal.

(1.15) P 366 Appendix: Could the list be completed with links to all models? A reference for each Earth system model would be helpful if one wants to learn more about the inner workings of the respective models.

We have added references to the description paper(s) for each model to Table 1.

(1.16) Technical remarks:
p360, l 17: . . . the the . . .

We have fixed this typo and thank the Referee for bringing it to our attention.

(1.17) additional citations:
Collins et al. (2011): Development and evaluation of an Earth-System model – HadGEM2, Geosci. Model Dev., 4, 1051-1075, doi:10.5194/gmd-4-1051-2011.
Giorgetta, M. A., et al. (2013), Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, J. Adv. Model. Earth Syst., 5, 572–597, doi:10.1002/jame.20038.
Hurrell et al., 2013: The Community Earth System Model: A Framework for Collaborative Research. Bull. Amer. Meteor. Soc., 94, 1339–1360. doi:http://dx.doi.org/10.1175/BAMS-D-12-00121.1

We would like to thank Referee #1 for the time and effort providing this helpful feedback on our manuscript.