Response to gmd-2016-114-RC1:

This is a well composed manuscript written in a lively style to describe experiments in finer grained component concurrency for at atmospheric model. The novel aspect is the use of component concurrency to improve scalability in an atmospheric code. In general the manuscript is appropriate for GMD, but revisions would strengthen the presentation. The suggested revisions are substantial, but of a nature that the editor can readily adjudicate.

2. I enjoyed the more lyrical style, but in places the style came across as glib, particularly when precision was sacrificed for poetry. Grounding the manuscript better in data and substance would strengthen the exposition greatly.

   We have taken some pains to sharpen the arguments and make them more rigorous at many places through the text, including the specific items highlighted by Reviewer 1 below. We hope we have retained some of the "lyricism" without giving in to "glibness".

3. The figures are substandard and not sufficiently quantitative. In particular differences between different configurations of the same simulation are preferred over differences to the observations, as the former, not the latter, is the point of the manuscript. Also attention to map projections and color scales is required.

   Please see reply to (25) below.

4. The frequent reference to Pauluis and Emanuel was insufficiently discriminating and in places misleading. The reference gives the impression that infrequent coupling to radiation is a substantial source of bias and also model instability. This is not what the manuscript is about. Moreover the Pauluis Emanuel study, while noteworthy, has not been shown to generalize. It might, but the literature is not there. In the grand scheme of things, the trade-off in accuracy and stability of calling radiation more frequently, versus simulating at higher resolution, is not well understood.

   We agree that that there need to be more studies of the effects of temporal subsampling of radiation. However, there is no case to be made that temporal subsampling is superior for *physical* reasons: at best, we can claim that we do not see marked increase in bias or RMS error as an effect of subsampling, and that it improves time to solution. The Morcrette papers we have cited in response to Reviewer 1's other comments (about spectral and spatial subsampling) also only claim expediency and decreased time to solution as a reason. They indicate error increases are tolerable over short timescales and probably increase model bias on climate timescales. Based on these results, we believe our focus on methods to possibly eliminate temporal subsampling are justified. We have modified our discussion of the Pauluis-Emanuel and Xu-Randall studies to indicate the need for further exploration of the effects of temporal subsampling, including solution convergence as Δt_{rad} and Δt_{atm} converge. See page 4, line 4.

5. Some of the hard-core computational issues are insufficiently addressed. In particular, component concurrency probably will affect the hard scaling floor, and the trade-off between communication vs processing in codes like ESMs with low arithmetic intensity. I guess there are also trade-offs that arise because of the need for the concurrency to be in a shared memory implementation. The manuscript would be strengthened if these issues were discussed in a more thorough manner.

While optimal performance for OpenMP threading is generally achieved when the threadpool is constrained to a single socket, scalability of concurrent components is limited by the number of cores in a shared-memory node. This number is slated to be very much larger than today, $\mathcal{O}(100)$ on MIC architectures. See also reply to (35) below.

6. There are too many acronyms, and some seem indiscriminately chosen.

We have gone through the text and noted the following acronyms: GPU, MIC, CCC, ENIAC, ESM, IPCC, FAR, AR5, HPC, FMS, SIMD, SPMD, MPMD, SMP, MPI, GFDL, OpenMP/OMP, AOGCM, LBL, PE, AM3/CM3, SST, AMIP, GPCP, CERES EBAF, NOAA, SYPD, CHSY, NPES, CPU. Most of these are quite well-known and commonly used, and cover standard terminology in the computing and climate literature. Some refer to institutions, models, and observational datasets. We hope none of them look "indiscriminate" in the revised text. All have been defined at first use, thanks to the close reading by reviewers: see (8) below.

7. The manuscript does not distinguish between AGCMs and ESMs, frequently discussing the results in terms of ESMs but then presenting results for the atmospheric GCM alone. Do the results generalize to problems that already have much more concurrency? I guess so but this is a separate point and the manuscript should discriminate between what was done and what is inferred based on what was done.

We have clarified our use of "ESM" as a generic term for models of any level in the complexity hierarchy. We have also more clearly explained that the current implementation is in an atmosphere-only setting, but readily generalized to more components, which we describe in our plans for future work. This includes components invoked only in ESMs defined in the narrower sense, that of models that include an interactive biosphere. See page 2, line 25; page 15, line 20; and the caption to Figure 2.

8. p1l2: Acronyms

Fixed, see page 1, line 2

9. p1l24: Is that true, or did the technology also drive things. it makes it sound like everything was possible and we just chose something, rather than necessity driving development.

It is mostly believed that the principal cause for the decline of vector computing was economics: commodity clusters were able to match performance of custom hardware at a vastly lower price. We have added some text and citations to bolster this argument, see page 2, line 9.

10. p2l5: I am not sure what figure the authors are trying to say here. Actually resolution has not kept pace with computing as far as I can tell, and the reference to the figure does not make sense.

There was an error: it was Fig 1.4 from the 2007 Report, not the 2013 one we cited. It is now fixed, and a link to the figure itself is provided. See page 2, line 14.

11. p2l7: Would help to explain to the reader the phrase "arithmetic and logic" is a memory fetch logic?

Fixed, see page 2, line 16

12. p2l11: “The state of play of climate computing in the face of these challenges” this phrase comes across as a bit of a throw away. Did the Balaji (2015) paper make a point that is important for the present discussion? If so what was it.

    Fixed, see page 2, line 23

13. p2l20: “and there is constant churn of operations” missing an article. . . lyrical, but it gives the idea that the computer is kept busy computing rather than moving information around. Most codes are memory bandwidth limited . . . which you get to shortly. But this intro sentence did not prepare me well.

    Fixed, see page 3, line 3

14. p2l30: Here: “Of the many factors of 10 increase in performance needed to get to the promised land of ’exascale computing’, we believe at least one can come from component organization.” I would prefer precision over poetry.

    Fixed, see page 3, line 16

15. p3l5: Weather centers also run spatially coarse-grained radiation. Also some have proposed a form of coarse graining in the spectral domain, i.e., Monte-Carlo Spectral Integration.

    This is an important oversight, and we have added a short discussion on spatial and spectral subsampling, see page 3, line 28. The main focus of this paper remains the elimination of temporal subsampling at modest computational expense.

16. p4l9: A reference is needed here. My intuition suggests that such high processor counts have only been applied to more traditional Atmosphere, or Ocean or Atmosphere Ocean Problems, but not simulations of the carbon cycle, i.e., the use of ESM here is misleading as earlier it implied biology.

    The reviewer is correct: most of the very high end PE counts are for specialized problems, not full ESMs in climate mode. We have clarified this, and added some references: see page 5, line 8

17. p6l15: Subscript abbreviations are usually written in roman font.

    Fixed, here and everywhere else in the text: see page 7, line 11.

18. Figure 3: I spent some time on this and I am not sure I understood it. The vertical dimension denotes sequence, first to last from top to bottom. The boxes indicated either the legend or a component process? The thickness of the box denotes? A figure should be illustrative, not a riddle. Also a bit more structure might help me understand what atmosphere up is. The meaning of some colors seems to be specific, others decorative?

    Noted. We have reworded the caption to Figure 3 to clarify these points.

19. p7 l2: “interact strongly with atmospheric chemical species and clouds” this could exclude water vapor, why not say, couple strongly to composition.

Text clarified: see page 9, line 6.

20. p7 l4: By this definition the ocean does not have tracers. Aren't tracers really just scalar quantities that are transported with the flow subject to source and sink processes.

    Text clarified: see page 9, line 7.

21. p7 [reviewer typo: this is p8] l6: Seconds and hours (line 9) are units, and can be abbreviated.

    Done: see page 9, line 11.

22. p7 [reviewer typo: this is p8] l9: This makes it sound worse than it may be, some codes rescale the radiative heating rates at each timestep by the insolation, or the surface temperature, in a sense linearizing about the state defined every 3 hrs. Something you mention on the next page, but it comes late.

    Agreed, eliminated last sentence of paragraph, as it's redundant with what come's next: see page 9, line 13.

23. p7 [reviewer typo: this is p8] l19: Spell out PE . . . Processing Elements? A socket?

    Fixed: see page 9, line 22.

24. p7 l14 [reviewer typo: this is p8 l24]: "Architected"? Okay, it can be used as a verb; but I think designed, or constructed would be better.

    Rewritten: see page 9, line 28.

25. Figure 5: This needs redrafting, first for the colour scale (no rainbows); second to show the common color scales in those panels where it is appropriate. The highly distorted projection should either be motivated or replaced.

    If the objection is to the use of the Mercator projection: we agree that it introduces distortions, particularly areal distortion as you approach the poles. While this is known, this projection is customary and in very widespread use in the literature. We do not believe the use in this instance to be egregious, as the results are in no way compromised by the projection.

    Similarly, we have used a somewhat standard color scheme: the rainbow in panel 1, and a warm/cool two-color palette for the difference plots. This again is quite customary.

    We are somewhat puzzled by this remark as neither the projection nor the color scheme are in any way misleading or distorting the results.

    That said, we have modified Fig. 5 and 6 to show the model-obs. differences in Panel (b), and the (smaller) model-model differences in Panels (c) and (d). We hope this makes the discussion clearer. See page 11, line 21, and captions to Figs. 5 and 6.

26. Figure 5: I would like also to see land temperatures. And differences among the simulations are far more interesting than differences with GPCP

Again, we are puzzled by the remark and unsure how (or why) to add land temperatures to a global plot of precipitation. No land or ocean cells have been masked on this plot.

The differences between the simulations are interesting, but cannot be over-interpreted, as in practice we would retune the models to maintain the same top-of-atmosphere net radiative flux. The focus of the paper is on the computational aspects and the coupling algorithm. We expect to visit this issue in greater detail in a future paper describing scientific model results from concurrent-radiation run where $\Delta t_{\mathrm{rad}} = \Delta t_{\mathrm{atm}}$, with appropriate tuning. See also response to (28) below.

27. P10l14: "Remarkably" similar. . . my rule of thumb is that people use the word remarkable when they don't know what of substance to remark.

Rephrased, see page 11, line 22.

28. p10l30: I would make the tuning point later, as it seems as though the authors are interpreting the differences as fundamental, rather then simply an illustration of compensating biases in a manner that is to be anticipated.

The tuning discussion comes at the tail-end of Section 4.1, which is the last discussion of physical results. We agree that the differences are not fundamental, and in the initial draft had indicated that they were "within the margins of the tuning process." We have expanded on that to make it clearer, as suggested, see page 14, line 2.

29. p11l5: SYPD has a unit, i.e., yr d^-1

SYPD is itself a unit, in fact it is introduced using the phrase "simulated years per day, or SYPD", see page 14, line 6. We believe it would be redundant to attach the unit again. Units are mandatory when the same quantity can be expressed in different units, e.g m and cm. That is not so in this case.

That said, we have added a discussion and definition of SYPD, see page 14, line 6.

30. Table 1: Maybe spell out the acronyms; for example does the introduction of CHSY really help anything? And if so why not PHSY, processor hours per simulated year. The computational cost of radiation appears small. If increasing the frequency of radiation nine-fold increases the computational cost by 50% this implies that the cost of radiation is about 5% of the total computational cost in the default configuration. Is this correct? If so this is rather small compared to some other models, suggesting that the proposed approach might be even more beneficial for other centers, or offer the possibility of more exact representations of radiative transfer. Here some clear numbers would be useful.

We do indeed believe the CHSY discussion is useful, as both time to solution (SYPD) *and* resource consumption (CHSY) are taken into account in configuring the parallel layout of a production model. See discussion in Balaji et al. (2016). We have also chosen to be consistent with that paper in naming the throughput measure CHSY instead of PHSY, which we agree would have also been a valid choice.

In regard the second point, the reviewer is correct, and we have made this point now in the text, see page 14, line 17.

31. p13l13: This paragraph is a bit ungrounded in the manuscript, which does not evaluate MPMD approaches. Certainly the GPU rewrite of COSMO has a factor of 3.6 speed up on a first implementation . . . so there is some room for efficiency gains through reprogramming, also the inexact hardware approaches (Dübben and Palmer) merit mention if this paragraph were to be retained and better grounded in the manuscript.

    It's true that the background material is a bit ungrounded in *this* manuscript, but summarizes the findings of Balaji (2015), which does indeed survey MPMD methods and inexact computing approaches. We have rewritten the paragraph to reflect that these are findings from Balaji (2015). See page 15, line 14.

32. p13l20: Where does the order ten components come from. I think "probably not more than ten" would be more accurate, but in either case if this comes at the end it should also be better ground in the manuscript.

    We have toned down the speculative statements here, see page 15, line 28.

33. p14l22: By data do the authors mean model output? Or the performance data, gleaned from the benchmarks? The use of "data" suggests the latter, but the former should also be addressed.

    Fixed, see page 16, line 18