Reply to reviewer comment RC2

The paper by Bishnoi et al. describes the work on using the modular supercomputing architecture (MSA) for coupled atmosphere and ocean simulations of the ICON model. The authors find that the MSA-approach improves the energy consumption by 59% compared to running the entire coupled model on the CPU nodes of the JUWELS Cluster supercomputer. The paper is well written and the results are very relevant for the audience of GMD. It is overall well structured and provides all the necessary details to understand how the results were obtained. There are a few places which I describe in detail in the specific comments below where the text is difficult to understand and should be improved. I recommend to publish the paper once these comments are addressed.

We thank reviewer 2 for the constructive comments and suggestions and for the positive impression our work has made. In the following we tried to carefully address your issues and hope that our corrections will result in a further improved paper manuscript. In addition, we added a short paragraph summarizing the discussion that arose from community comment CC1.

During the preprint phase, we detected an error in our Thermal Design Power (TDP) calculation. In fact, the TDP reported for JUWELS Booster was 400 W per CPU instead of 250 W. Taking this into account, our estimates for energy saving by using MSA instead of JUWELS Cluster reduced from 59% to 45%. We are convinced that also this somewhat smaller improvement justifies publication and hope that the rough estimate can be replaced by more accurate measurements in the future.

Specific comments:

line 33-35: This sentence is difficult to understand since it is too long. It should be split in two: "The code of more advanced complex climate models is composed of many different kinds of operations, e.g., ... Another level of complexity is added through optimisations for specific computing hardware (Lawence et al., 2018)."

done

line 62: up to 10% overall speedup according to line 360 doesn't sound satisfactory

In line 360, we refer to potential additional energy savings when porting ICON-O to GPUs, not speedup. We changed “savings” to “energy savings” in line 360 (now line 358) for clarification. Work on GPU porting of ICON-O is ongoing, aiming at better standalone performance and, in coupled mode, reducing parts of the coupling overhead (data transfers CPU host – GPU device).

line 83-87: Mixing sections and subsections is confusing. Better write something like: "In Section 2 we provide a comprehensive description of the ICON model and its specific setup. Section 3 presents a brief overview of the MSA, with an introduction to the concept (Section 3.1), a description of the modular hardware and software architecture of the JUWELS system at JSC (Section 3.2) and the strategy for porting the ICON model to the MSA (Section 3.3). Section 4 contains the results from our analyses for finding a sweet spot configuration for ICON (Section 4.1), the comparison to a non-modular setup (Section 4.2), and strong scaling tests (Section 4.3)."

Lines 83-89 have been rewritten according to the suggestions of both reviewers.

line 225-241: The goal of these two paragraphs should be made clearer. In particular the formulation "until we reached user allocation limits" in line 237 sounds to me as if ICON-A was still slower compared to ICON-O and more nodes beyond the allocation limit would have improved the energy efficiency of the homogeneous configuraton further. This is not the case according to line 295 *which states that ICON-O and ICON-A are balanced. Also it should be mentioned that this will be described in much more detail in Section 4. I suggest to replace the entire two paragraphs (line 225-241) with something like: "To quantify the benefit of the MSA approach we compare the energy consumption of an optimal MSA configuration with a homogeneous setting in which the entire coupled model is run on the same type of nodes while keeping the run-time roughly the same. Since not all model components of ICON can take advantage of GPUs we use the CPU nodes of the JUWELS Cluster module as a baseline for this comparison. The run-time is kept roughly the same by using the same number of nodes for the ocean component. Both configurations (MSA and homogeneous baseline) are optimized by adjusting the number of nodes used for the atmosphere component such that waiting times between model components are minimized. All other model parameters are kept the same. The process of finding the optimal configuration is described in detail in Section 4.1 for the MSA configuration and in Section 4.2 for the homogeneous baseline. In addition, we performed a strong scaling experiment to prove the scalability of the MSA approach which is presented in Section 4.3."*

The last paragraph of Section 3.3. has been re-arranged according to your suggestions.

Figure 3: The absolute values of the coupling time alone are not very meaningful in this context. Either mention already here how these times compare to the overall run-times or use percentage of the overall run-time of the simulation for the horizontal axis. This would allow the reader to immediately understand the severeness of these coupling times and it would still convey the message which configuration is best.

We took up the reviewer's valuable suggestion and changed Fig. 3, showing now the percentage of runtime instead of absolute timings.

line 269: If you don't show the percentage of the overall run-time in Figure 3 you should menion here the overall run-time to allow the reader to understand why it is a significant portion of the overall run-time.

done with Fig. 3

line 285: I love the way how you compare the MSA and non-MSA approach. I agree that it is a fair comparison. I just miss a clear statement how you chose to compare the two approaches. In my opinion keeping the number of ICON-O nodes the same is rather a matter of how you chose to compare the approaches than a matter of making the comparison fair. In principle one could also choose to compare how many SDPD the same amount of energy can achieve in which case one wouldn't keep the number of ICON-O nodes the same. I think you should replace "In order to make a fair comparison" with something like "In order to keep the run-time roughly the same"

We agree that a comparison could also be done in different ways. We changed "In order to make a fair comparison …" to "We decided to scale up the non-modular simulation such that total run times are in the same order than for the MSA setup. To achieve this, …" for clarification.

line 295: What exactly does it mean that atmosphere and ocean are balanced? Does it mean that coupling times are minimized or that the integration times are the same?

We increased the number of GPU nodes used for ICON-A such that integration times for atmosphere and ocean are roughly the same. We changed "workload for atmosphere and ocean" to "workload for the integration of atmosphere and ocean" for clarification.

Table 3: Please explain in the caption or in the main text whether the waiting time is already included in the total time or not.

The total timers includes the waiting times both for the halo exchange and the waiting for the exchange of coupling fields. We added an explanatory sentence in the caption of Table 3.

Figure 4: What is the relation between the numbers in Table 3 and the results shown in Figure 4? The figure shows a speedup of about 1.8x for atmosphere coupled at 355 nodes. All the speedups for the atmosphere component in Table 3 give me a speedup of at least 2 times even if I assume that the waiting time is not included in the total time. How does this fit? Why is the speedup at 355 nodes for the atmosphere lower than at 237 nodes? This is not visible in Table 3.

Thank you for pointing to this inconsistency, which has also been noticed by reviewer 1. The numbers given in Table 3 are correct, but we used erroneous values for producing Fig. 4. Indeed, speedup for 355 nodes is well above two and speedup increases up to this node count. Figure 4 has been corrected in the final version of the manuscript.

Minor comments:

line 13: "combination 84 GPU nodes" => "combination of 84 GPU nodes"

done

line 42: "factor of 1 million" => "factor of more than 1 million"

done

line 91: "non-hydrostatic atmosphere" => "non-hydrostatic atmosphere model"

done

line 96: It would be good to introduce the R2B9 grid already at this point which is often used later.

done

line 117: It would be helpful to mention already here that all experiments run for 1 simulation day.

done

second line of the caption of Figure 3: "Cluster / Booster" => "Booster / Cluster"

done

third and fourth line of the caption of Figure 3: "The number of Cluster nodes (17) dedicated to I/O is not taken into account, since it is kept constant across all experiments." => "The 17 Cluster nodes dedicated to I/O are not taken into account, since the number of IO nodes is kept constant across all experiments."

done

line 268: "(48)" => "(48 cores/node)"

done

line 269: "(85/80)" => "85 Booster nodes / 80 Cluster nodes"

done

line 295: "ICON-O" => "ICON-A"

done

line 296: "ICON-O" => "ICON-A"

done

caption of Table 2: Shouldn't this be rather "MSA configuration" than "MSA architecture"?

done

line 358: "simulationd" => "simulations"?

done

line 401: please remove this line "Hallo Olaf, wir testen jetzt...."

done