**We thank the reviewer for the comments. Before we answer the comments, we need to point out that we found an error in model runs on the Prometheus cluster. The error caused collision-coalescence to be not modeled and made condensation time step to be 0.5 s. In typical cloud simulations collision-coalescence is modeled and condensation time step is around 0.1 s. The error significantly reduced time complexity of microphysics computations. The error has been fixed and simulations were repeated. There are important differences between the faulty and corrected Prometheus results. Discussion of results, summary and abstract have been reworked to account for these differences.**

**Answers to the reviewer comments are listed below.**

General Comments:

The paper describes a port to GPU for the University of Warsaw Lagrangian Cloud Model (UWLCM). The authors have identified a sensible separation, handling Eulerian terms (bulk flow) with CPU computing and offloading the Lagrangian terms (particle) to the GPU. Alongside the effort involved in rewriting the code for GPU, the careful splitting of terms and balancing the workload between CPU and GPU are this paper's main original contribution.

Section 1 gives a good overview of cloud modelling  certainly suitable for those unfamiliar with the topic. Combined with section 2 this gives a brief but sufficient overview of some cloud modelling literature and UWLCM. Section 3 gives a good description of the authors contributions in terms of code modification strategy.The breakdown in section 4 describing simulation hardware and the performance results is good and the level of detail is suitable for most of the analysis.

The results look very promising, but would benefit from some significant changes to the presentation and corresponding modification in the discussion in the text. (See specific comments). Reproducibility is also considered, the authors' code is available, the system stack is fully specified as well as hardware used.

Specific Comments:

There are some key areas where changes need to be made:

1. One of the key contributions of this paper is the author's careful handling of balance between CPU and GPU workload. Specifically, trying to simultaneously utilise both the CPU and GPU as much as possible. However, the treatment of this in both the text and figures needs further clarification.

In the abstract, this was unclear and this reviewer thought it was a mistake, a rewording here would clarify the term. The text should also have a definition of what the authors intended with the term "amount of parallelisation".

This lack of clear definition is compounded by Figures 2 and 4. These figures should be modified to present the information more clearly.

As I understand the total wall time per timestep (t_tot) comprises of:

t_tot = t_both + t_CPU + t_GPU

Where

- t_both is the data presented in blue (CPU + GPU) the time that the authors are aiming to maximise to increase the "amount of parallelisation".

- t_CPU is time where only the CPU is performing work and the GPU is idle

- t_GPU is time where only the GPU is performing work and the CPU is idle

**That is correct. We modified the abstract so that it does not include the term “amount of parallelisation”. A new subsection “Performance metrics” has been added to the manuscript. In it we define t_tot, t_CPU, t_GPU, t_CPU&GPU and what we mean by “parallelization”. In Figs. 2 and 4 ambiguous terms have been replaced with t_tot, t_CPU, t_GPU or t_CPU&GPU.**

My suggestion for improving the figure is having these quantities stacked, rather than overlapping, either as a shaded chart or preferably a stacked bar chart. This would more clearly present the t_tot quantity of interest, which is currently hidden. The structure of the plot should also be outlined in the text.

If any of my assumptions here are incorrect, there should be clarification in the text as to what is actually being plotted.

**Values of t_CPU, t_GPU and t_both are stacked in Figures 2 and 4. This has been clarified in figure captions. We use a stacked shaded chart instead of a stacked bar chart, because values on the horizontal axis are not uniformly distributed and the shaded chart is more readable.**

I am not opposed to the auxiliary plot in red, but in Figure 2 the y-limits should be fixed as in Figure 4 so comparisons between the plots can be made.

**The y-limits have been fixed as suggested.**

2. In two places in the text, power and energy are mentioned but not discussed. Either, these comments should be removed as there is no backup or discussion of the assertions. Or, power information for both CPUs and GPUs needs to be presented in table 1 and at least an estimation of energy/power usage needs to be presented and a comparison between CPU vs CPU+GPU made.

**Information about thermal desing power (TDP) of CPUs and GPUs has been added to table 1. Estimated energy used in CPU and CPU+GPU is now plotted in Fig. 3.**

3. The discussion from lines 150-156 along with Figure 3 doesn't really make sense. CPUs and GPUs cannot be compared in such a way, and the assertion that a certain number of CPU cores are in any way equivalent to a GPU is misleading, since the basis of comparison is wall time only. This doesn't take into account aspects like power consumption or accuracy, for instance. From the discussion, the quantity of interest here is total time (per timestep perhaps), this quantity should be plotted for both CPU only and CPU+GPU case so readers can compare. Similar figures could be

used to compare the energy/power usage for CPU vs CPU+GPU and the accuracy, should the authors desire to draw a comparison.

**The CPU and GPU comparison has been changed. In Fig. 3, wall time and energy use per time step are now compared, instead of the hypothetical number of CPUs that would replace GPUs. The discussion has been changed accrodingly.**

Whilst the changes to the plots may be significant, I believe that they can be made without performing additional simulations, just by changing plotting scripts. However, I believe that the changes should be further reviewed so I have suggested major, rather than minor revisions.

More minor corrections, line by line notes:

L4: (Abstract) Amount of parallelization, needs clarification. *See 1 above

**Corrected, see answers above.**

L10: Moore's Law hasn't really ended, Dennard scaling has, I can't take too much issue here though, as the cited paper backs up your point.

**This has been replaced with:**

**“As CPU clock frequencies no longer stably increase over time and the cost per transistor increases, (...)”**

L89: Some expected speeds should be given for expected PCI express and interconnect speeds, for comparison (could be added to table)

**PCI-E and interconnect speeds are now given in table 1.**

L90-91: Clarification, what technology is allowing GPU-GPU inter-node communication? This should be mentioned.

**We added:**

**“Intra-node communication between GPUs controlled by a single process makes use of the NVIDIA GPUDirect Peer to Peer technology, which allows direct transfers between memories of different devices. Communication between GPUs controlled by different processes is handled by the MPI implementation. If the MPI implementation uses the NVIDIA GPUDirect Remote Direct Memory Access technology, inter-node GPU-to-GPU transfers go directly from device memory to the interconnect, without host memory buffers.”**

Table 1: Please add memory bandwidth figures for both main memory (RAM) and GPU memory as well as theoretical peak flops for CPUs and GPUs.

**Done**

L156: Power usage is mentioned here with no prior discussion or further explanation. Either remove comment about power, or include wattage for CPU + GPU in table 1 and discuss this fully. *See 2 above

**See answer to comment 2.**

L217: Energy here is again a throw away comment. *See 2 above

**See answer to comment 2.**

L223: "A simulation with 20 million grid cells and 2 billion particles" I like the discussion of the number of DOFs here, the earlier text and possibly also figures would also benefit from some concrete discussion of the problem size. (I'm aware of table 3, but would be useful to have some discussion in text)

**We added:**
**“Depending on the scenario, number of Eulerian grid cells is between 0.5 and 18.5 million, and number of Lagrangian particles is between 40 million and 18.5 billion.”**
**to the discussion of testing scenarios in section 4.5.**

Figure 5: Would benefit from a line indicating the ideal scaling in each case.

**Such line has been added.**

Technical Corrections:

**All suggested technical corrections were made.**

I have ignored any British English vs American English discrepancies.

What follows are my interpretations/suggestions for typos and grammar:

Parallell -> Parallel (throughout)

L18-19: what is known as cloud parameterization -> which is known as cloud parameterization

... and throughout "what" -> "which" L22 for example (what requires) L27, L65, L183

L31 gained in -> gained

L37 Main goals -> The main goals

L71 parallely to -> in parallel with

L79 Strategy -> The strategy

Table 2: List of software of servers used -> List of software on servers used. (I think this is what was meant)

L129 ran -> run ... and throughout L134,L159,L178

L209: parallelly done -> done in parallel

**Additional changes**

**We fixed an error in the link to the UWLCM code.**