The authors described the activity of porting the ICON-A atmospheric model to a GPU-based parallel architecture using a directive based approach for the parallelisation in order to take full advantage of exascale architectures and improve scientific outcomes resolving physical and climate process down to the scale of a few kilometers. First of all, I would like to express my full appreciation for a really well written manuscript and for a wealth of information and details.

However some points can be better clarified / discussed

- The GPU approach followed by the authors leverages on OpenACC, although OpenMP is mentioned in some parts of the manuscript (lines 499, 833). My questions are:
  + is the initial version of ICON-A parallelized with MPI + OpenMP? If so, it should be explicitly mentioned at the beginning when the description of the model is given.
  + taking into account that OpenMP v5 supports the GPU offloading, supporting not only NVIDIA but also Intel GPUs, why did the authors choose to use OpenACC instead OpenMP? And finally, once the choice fell to OpenACC why the OpenMP is kept inside the code? The authors should clarify these aspects to better justify their choices.

- in Section 4.4 Line 520 the authors first introduce the concept of reproducibility which will be better discussed later in Sections 5. In Section 4.4 the meaning of the word "reproducibility" is not clear. Are the authors referring to the bit-identity reproducibility or tolerance-reproducibility? How was reproducibility evaluated in the context of physical parametrization (Sec 4.4)? Moreover, the authors uses the "ACC LOOP SEQ" directive to "fix the order of the summands" but it is not clear why this is needed; what is the correct order to do a summation. Considering that the round-off error is inherently present in the code, even in the sequential version of the code, why should summation follow the "LOOP SEQ" order?

- In Section 5 the authors deeply discussed the validation techniques available for ICON. Namely, in Sect 5.3 the tolerance testing is presented, which consists of evaluating an
ensemble obtained by perturbing the state variables with a uniform error of the order of magnitude $10^{-14}$. My comment here is that the main source of divergence in the outputs, when implementing a parallel version of a code, is due to the round-off error that can grow after several time steps. In order to evaluate the effect and impact of the round-off error it is probably best to create an ensemble by changing the order in which the grid cells are evaluated, such a by shuffling the arrays with the grid cells.

- In Section 6.5.1 the authors should provide a comment on why the radiation exhibits a super linear strong scalability on PizDaint and not on Juwels-Booster neither on Levante. Why the transport and the vertical diffusion have a super liner strong scalability on Levante? The vertical diffusion has a really strange and counterintuitive behaviour since its scalability curve increases with the number of nodes.

- In Section 4.3.1 Line 322 the sentence "There are code divergences in the non-hydrostatic solver" is a bit misleading since it is not clear whether it refers to thread divergence or code differences between CPU and GPU.

- Listing 2 reports an example to explain the use of scalars on GPUs instead of arrays, but the transformation of 2D array into a scalar is not fully clear; namely, in the expression for the scalar (line 331-333) the index $j_{k-1}$ is used while in the expression (lines 336-338) the index $j_k$ is used. Moreover is also unclear whether the "$z_w_{concorr\_mc\_m1}$ values are used/needed after the do loop; if these values are not used outside the loop probably the scalar transformation is also useful for the CPU case.

- In the abstract and in conclusions the authors write that the model exhibits a good weak scalability. But after a careful reading and according on what is stated in Section 6.5, "ICON exhibits very good weak-scaling for a 16-fold increase in node count", actually, a complete weak scalability analysis has not been provided as the weak scalability has been evaluated only in the case of 16-fold increase. I suggest that the same comment is also report in the abstract and conclusion.

More cosmetic comments, suggestions and typos:

- In the abstract, line 8, it is better to use "kilometres" instead of "km"

- Line 17-18: there is a pun in the sentence... the weak scalability is good and the strong scalability is weak

- Line 116: "(black)" should be "(blue)"
- Line 272: "data types" should probably be "data structures"

- Listing 6 and Listing 7 have exactly the same caption. I suggest to merge together both listing or to differentiate the captions

- Line 702: "ptest" mode is mentioned here for the first time, it would help if, in the same sentence, the authors anticipate that the mode is described in the following section.

- Line 787: "Ss and Ss" should be "Ss and Sw"

- Line 919: "1 SDPD" should probably be "1 SYPD"