General comments:

The paper provides a good description of a single and mixed precision approach in atmosphere modeling, with adequate motivations for current and future models. The authors describe their strategy for building the single and mixed precision models, as well as the justification for their choices. The results are divided in a computational performance and a physical performance sections, which help the reader understanding the benefits/disadvantages of each approach both from a computational perspective as well as from a model skill one.

Overall, this is a good paper, and I recommend its publication. However, I would like to suggest a few aspects that the authors could clarify better, or provide some perspective on.

Reply:

We would like to express our gratitude to Dr. Luca Bertagna for detailed review and constructive feedback on our manuscript. We appreciate your insightful comments and suggestions. Each issue you highlighted has been thoroughly addressed in our revised manuscript. Your expertise has been essential in enhancing both the quality and clarity of our work.

Specific comments:

1. From the text in section 2.2, it seems the authors adopted a "greedy" approach, where each choice of variable to treat in single precision is built on top of the previous. Is that the case? Have they tried to start from different choices for the initial single precision variable? If so, what did they observe? If not, do they expect differences? I'm thinking that all paths would led to a similar final configuration of variables to be switched to single precision, but would be nice to know for sure if that's the case or if different initial choices may lead to different final configurations.

Reply:

Yes, a greedy-like approach was employed to identify precision-sensitive variables, where the selection of each single-precision variable builds upon the preceding one. The computational sequence of the code mirrors that of the double-precision version. Initially, in selecting single-precision variables, we systematically attempted to reduce the precision of variables encountered sequentially in the code, starting with the first variable, followed by the second, third, and so forth. We did not explore alternative initial choices for single-precision variables.

Overall, we are confident that various paths would converge to comparable final configurations of single-precision variables. As detailed in Section 2.2, the precision sensitivity varies among different computational terms, with some being highly sensitive while others are less so. This suggests that precision sensitivity is inherent to specific terms and largely unaffected by the chosen path.

Nonetheless, we acknowledge the potential for further optimization and time reduction.

2. The authors say that "the precision optimization tests were conducted using the G8 grid". I assume the resulting configuration of single precision variables was also used for the G6/G7 tests in the section 4.2. Have the authors confirmed that the same configuration was indeed optimal (by means of adding/removing some SGL

variables) also for other grids? If not, would they expect any difference? Why or why not?

Reply:

The initial development of the mixed-precision code was mainly based on the G8 grid and the Jablonowski-Williamson (dry) baroclinic wave test, as described in Section 2.2. Subsequently, this code was utilized for all tests presented in this paper, encompassing both higher and lower resolution grids.

While we cannot guarantee that the optimization results are universally optimal across all grids, existing literature and our own experiments suggest that the deviation between reduced-precision codes and double-precision codes typically tends to amplify with higher resolutions. Therefore, we tentatively hypothesize that the mixed-precision model derived from the G8 grid demonstrates relatively minor deviations from the double-precision model on coarser grids. Additionally, the current 5km test appears reasonable for higher-resolution grids.

In the future, we aim to further reduce the precision of certain variables and conduct more tests at the kilometer-scale to ensure the robustness of the reduced-precision model code.

3. When formatting equations in section 2.3, be aware of color-blind people. I am not one, so I can't give a thumbs up/down. But I would suggest to verify that the color choices are not affecting color-blind readers. E.g., the blue term in eq 10 may not look blue, or the green term in eq 5 may not stand out from the red around it. As a possible alternative, you could consider underlying, or use another box (perhaps dashed, to distinguish from the black one).

Reply:

In response, we have revised the presentation of the equations in Section 2.3, utilizing symbol annotations instead of color markings for clarity. Additionally, we have compiled the equations alongside the code modifications in an illustration (Fig. 1).

4. In section 3, the authors mention that SGL, MIX, and DBL all used the same computational resources. I would assume that, among other things, a reduced precision could allow to use less computational resources, which would additionally benefit performance (due to reduced MPI costs). This can be particularly beneficial on machines with sub-optimal interconnect, as well as on GPU architectures, where increased computational intensity (in terms of degrees-of-freedom per GPU) can increase overall performance. Have they explored this avenue? What are their thoughts on this?

Reply:

We agree "a reduced precision could allow to use less computational resources, which would additionally benefit performance (due to reduced MPI costs)." This is indeed the case. With fewer computational resources, the memory usage per processor increases, exacerbating memory bottlenecks. The use of mixed precision can alleviate the memory bandwidth bottleneck.

While not pursued in this paper, there is another work that has ported the mixed-precision code of this study to the next-generation Sunway computer. The processor of Sunway has 6 core groups (CGs), each of which consists of one management processing

element (MPE) and 64 computing processing elements (CPEs) organized as an 8×8 array (390 cores per processor). Those tests were conducted at 1-3 km horizontal resolutions. This may correspond to the GPU-like situation you mentioned, where "increased computational intensity can increase overall performance".

A notable observation is that mixed precision typically does not yield significant speedup on the MPE side but provides notable speedup on CPE-parallelized kernels. Considering that the Sunway architecture generally does not exhibit higher calculation performance in single precision compared to double precision, except for division and elemental functions, we can infer from the results that the MPE code is computation-bound. On CPEs, mixed-precision code demonstrates better speedup. One possible reason is that the CPE code appears to be constrained by memory bandwidth, and mixed precision reduces data size, conserving memory bandwidth and increasing cache hit ratio.

5. On line 182: why is 24% in parentheses? Seems like the line should be "27%, 24%, and 44%".

Reply:

Thank you. This was an error, and we have corrected it in the revised manuscript as "27%, 24%, and 44%".

6. Fig 2 seems to show absolute L2 error. It may be helpful to show a relative error, so that the reader can better gauge the impact of the SGL/MIX approximations.

Reply:

We apologize for not clearly defining the error norms in the initial manuscript. We have added an appendix in the revised manuscript to explain the error definition. Each error norm is now normalized by the true value of each variable, so the errors in Figure 3 (initially Figure 2 in the submitted manuscript) are relative errors.

Additionally, we have changed the errors in Figure 2 (initially Figure 1 in the submitted manuscript) from absolute to relative errors to present the results more intuitively. Correspondingly, we have revised the text to reflect this change.

7. Have the authors tried to see whether, for a fixed set of SGL variables, the quality of the SGL/MIX approximations (compared to DBL) changes with respect to numerical choices (such as the order of numerical schemes)? If not, do they expect similar quality?

Reply:

In response to this question, we have conducted additional experiments. We adopted the passive transport test with a Hadley-like meridional circulation (Kent et al. 2013) to specifically investigate the impact of reduced precision on various nominal-order advective flux operators.

We tested a range of icosahedral grid resolutions (G5-G9, approximately 240 km to 15 km) and employed different nominal-order (2nd, 3rd, 4th) horizontal advective flux operators. The results indicate that, regardless of the chosen horizontal flux operator, the single-precision simulations are largely comparable to the double-precision simulations across all resolutions, exhibiting nearly identical convergence rates. The findings from this test support our optimized results, indicating that the advective components of the equations are not significantly sensitive to the precision level, irrespective of whether higher or lower-order operators are utilized.

It would be nice to know if the need for specific terms in double precision comes from the underlying physics and PDEs, rather than from the particular details of the numerical scheme.

The complexity of determining the necessity for double-precision terms spans across multiple levels: from the fundamental physical laws yielding the raw PDEs, to a specific model formulation (like GRIST), and finally to the particular code implementation of that model (like this version). So it is not easy to clearly say which one is determining the outcome.

Safely speaking, our perspective is that the current optimization outcome is primarily relevant to the model and code under consideration and may not be universally applicable across all models. This assertion is easy to verify because there are already several dynamical cores using nearly-pure single precision (e.g., Váňa et al. 2016; Nakano et al. 2018) without performance loss.

On the other hand, the outcome implies that, within the current implementation, the advective part of the model demonstrates greater resilience when subjected to changes in precision. We may also argue that one can more confidently adopt reduced-precision for the advective part, while the pressure-related and some other terms may require more careful code implementation. This is an inference but can be a reference for other models.

The relative vulnerability of the pressure-related term may have some potential sources. First, pressure-gradient terms tend to suffer from the cancellation of significant digits, as also outlined by Nakano et al. (2018). Meanwhile, pressure-related terms are more tightly related to the fast processes in the atmospheric dynamics, and thus they could potentially amplify undesiable small pertubations. Advective process is relatively slower.

Looking ahead, we aspire for a further optimization and code modifications to achieve more reduced runtime while preserving the robustness of the physical performance (e.g., stability, accuracy, convergence).

8. While single precision is definitely more appealing at km-scale, it can still be interesting to use it at coarser resolutions. For instance, it could allow running larger ensembles, benefiting UQ investigations. The authors mention the G6, G7, and G8 grids, which are all km or sub-km grids. Have they done any experiment at lower resolutions? If so, did they observe similar patterns? If they haven't done such experiment, are they planning to? Why or why not?

Reply:

This question is referring to the splitting supercell thunderstorms test in Section 4.2. In this particular experiment, which was conducted on a reduced Earth radius, we employed the G6, G7, and G8 grids, corresponding to horizontal resolutions of approximately 1 km, 0.5 km, and 0.25 km, respectively.

To further explore the effects of resolution on precision, we incorporated additional tests at coarser resolutions using the G4 grid (~4 km), and detailed descriptions of these tests have been added to the revised Section 4.2. At this coarser 4km resolution, we observed that the differences between mixed-precision and double-precision simulations were significantly smaller than those noted in the km or sub-km scale

simulations, aligning with earlier experiences.

From 1 km to higher sub-km resolutions, the differences do not monotonically increase with finer resolutions. This phenomenon likely stems from the design of the DCMIP2016 supercell test, which is intended to ensure converged solutions as resolution increases (Klemp et al. 2015). The diminished precision sensitivity at 0.25km is likely related to this convergence.

Please note that in the original manuscript, the horizontal section shown in Figure 3 was not actually interpolated to a height of 5 km. We have corrected this mistake in the revised manuscript.

9. In terms of reproducibility, it would help if the authors could share a snapshot of the source code repo, containing all the needed modifications. It would also help to share (perhaps in the form of README files in that same repo) instructions on how to run the particular experiments they ran (e.g., input files, run scripts, peculiar environment settings,...).

Reply:

In the revised manuscript, Figure 1 now illustrates how the code has been adapted to facilitate mixed-precision computing. Additionally, we've included Section 2.4, which offers a concise overview of the major modifications made to the original code repo.

For all the test cases addressed here, the corresponding code, build and runtime configurations, as well as necessary input data have been made publicly available on Zenodo. The readers can replicate these tests within their own local environments.

Again, it is emphasized that the computational savings reported may vary based on different configurations, such as grid resolution, computing environment, and compiler settings, among others.

Reference:

Kent, J., P. A. Ullrich, and C. Jablonowski, (2013), Dynamical core model intercomparison project: Tracer transport test cases. *Quarterly Journal of the Royal Meteorological Society*, 140(681)**,** 1279-1293.doi:10.1002/qj.2208.

Klemp, J. B., W. C. Skamarock, and S. H. Park, (2015), Idealized global nonhydrostatic atmospheric test cases on a reduced-radius sphere. *Journal of Advances in Modeling Earth Systems*, 7(3)**,** 1155-1177.doi:10.1002/2015MS000435.

Nakano, M., H. Yashiro, C. Kodama, and H. Tomita, (2018), Single Precision in the Dynamical Core of a Nonhydrostatic Global Atmospheric Model: Evaluation Using a Baroclinic Wave Test Case. *Monthly Weather Review*, 146(2)**,** 409-416.doi:10.1175/MWR-D-17-0257.1.

Váňa, F., P. Düben, S. Lang, T. Palmer, M. Leutbecher, D. Salmond, and G. Carver, (2016), Single Precision in Weather Forecasting Models: An Evaluation with the IFS. *Monthly Weather Review*, 145(2)**,** 495-502.doi:10.1175/MWR-D-16-0228.1.