The authors present an idealized test case (supercell storm) for global atmospheric models and briefly discuss the results from a large set of models that participated in a model comparison project (DCMIP2016). The premise of the study is good and useful. Demonstrating variability in the simulation of atmospheric phenomena due to differences in the dynamic core is an important topic and one that is often overlooked by the science community. The methods, analysis and even the figures are almost identical to a paper by Klemp et al. (2015), so the paper adds nothing new in this respect. The usefulness of the paper comes from the presentation of results from a large and diverse set of models. Despite this, the paper needs significant revisions to improve the presentation and additional analysis/simulations are needed to make the results clearer.

Thank you for your thorough review. We agree that the main thrust of this manuscript is to document a set of well-vetted solutions from many world-class dynamical cores. This not only provides a snapshot of model status at the time of DCMIP2016, but also acts as an inventory for which future modeling endeavours can compare their high-resolution solutions against; both as a sanity check as well as hypothesizing dynamical core impacts on non-hydrostatic phenomena.

The introduction was disappointing. No literature review is done on the impacts of dynamic cores on atmospheric problems such as extreme weather events. There are recently published papers on this topic that should be cited. In addition, more discussion is needed on the motivation for the study. Why do we need to compare dynamic cores? Do we expect significant differences and if so, do these differences explain some of the uncertainty observed in forecasts of extreme weather?

We have added some additional citations regarding recent work tying dynamical cores to extreme weather. That said, we would be happy to include discussion of additional literature that we may have overlooked.

We have also added some discussion of DCMIP's motivation and why it matters at 'the end of the pipeline' with respect to weather and climate simulations.

The results of the resolution sensitivity study don't appear to be converged at 0.5 km for several models. For example, there are significant differences in one or more fields going from 1 km to 0.5 km resolution for the following models: CSU, NICAM, ICON, GEM, TEMPEST. The authors should show a 0.25 km resolution simulation for one or two of these models to demonstrate better convergence properties. In addition, shouldn't we expect that as the grid spacing becomes very small, the model solutions and bulk statistics should be fairly similar across models? This is not the case with the current results so they are likely far from converged. Some discussion and additional example tests (as described above) are needed.

We would expect intramodel model solutions and bulk statistics to begin the converge at the higher resolutions in this study. That said, one of the obvious findings of DCMIP2016 is that non-hydrostatic phenomena at the resolutions tested here remain sensitive to numerical scheme, diffusion, and physics-dynamics coupling in a way that large-scale features such as baroclinic instabilities are not [Jablonowski et al., 2016].

While additional convergence tests may be enlightening, it is also computationally (and logistically) burdensome to complete new simulations at 250m resolution. That said, we agree that the reviewer's concerns are valid and have made care to note that these are targets for additional simulations either at the individual modeling center level or for future DCMIP projects.

Page 1, line 8; What is meant by "physics-dynamics coupling"?

This is the technical coupling between the dynamical core and representations of subgrid processes; commonly referred to as 'physics.' [Gross et al., 2016, 2018].

Gross et al. [2018] defines *physics-dynamics coupling* (PDC) as '... bringing together all the various discretized components to create a coherent model will be referred to here as physics–dynamics coupling. The term physics–dynamics coupling has evolved from the fact that the resolved fluid dynamics components are commonly known as the dynamical cores or simply "dynamics," and the physical parameterizations that represent the unresolved and underresolved processes and the nonfluid dynamical processes are collectively referred to as "physics."'

Page 1, line 11; What is the difference between convective-permitting and convective-allowing? These mean the same thing to me.

There is a bit of a 'gray' area in the definition of this by various modelers. Colloquially, a model 'permits' the simulation of a phenomenon if it can be discerned, regardless of whether it is under-resolved or not. A model 'allows' a phenomenon if it can be discerned and is credibly resolved. That said, we understand now where this can be confusing and isn't critical within the abstract so we have chosen to just include the term 'convective-allowing' especially since regimes pertaining to resolved convection in the atmosphere are a continuum and not cut discretely.

Page 2, sentence starting with "It is based on the work of. . ."; Sentence doesn't read well and needs a re-write.

As requested, this passage has been rewritten as follows 'This test is based on the work of Klemp and Wilhelmson [1978] and Klemp et al. [2015] and assesses the performance of global models at extremely high spatial resolution. It has recently been used in the development of next-generation numerical weather prediction systems [Ji and Toepfer, 2016].'

Page 2, line 15; Need few sentences on what a "reduced-radius sphere" is and what it intends to represent. Is the use of a reduced radius sphere a good approximation to true radius, global dynamics? Why do you need to simulate a supercell to study the performance of a global model? Wouldn't a global phenomenon (with non-hydrostatic processes) be better suited for studying the performance of a global model? How many horizontal grid points are used in the reduced radius simulations?

Using a reduced radius sphere allows for computationally-efficient simulations of O(1km) grid spacings in global models without modifying the numerical framework. Simulating supercells are important for non-hydrostatic development because A) the storms strongly stress non-hydrostatic numerics, B) they represent key atmospheric phenomena with high societal relevance, making them of importance to both weather and climate modelers, and C) they are currently unresolved in most global numerical modeling frameworks but that is projected to change over the coming decade or two.

We have added 'Reducing the model's planetary radius allows for fine grid spacing to be achieved without the added computational expense associated with adding grid cells a standard global mesh in order to achieve non-hydrostatic resolutions [Kuang et al., 2005]. Wedi and Smolarkiewicz [2009] provide a detailed overview of the reduced-radius framework for testing global models.' to the text to shed light on this approach.

Section 2.1; The notation in this section is confusing. What is Ueq a function of , z? If Ueq is the zonal velocity at the equator, then isn't u = Ueq? I also don't understand the transition to defining ubar after u. This whole section needs to be described better.

This notation has been modified per the suggestions of both Reviewers #1 and #2.

Page 3, line 1; This can't be gradient wind balance because no Coriolis force is shown in the equation.

Our apologies for this confusion. Technically, this is more akin to cyclostrophic balance. The text has been clarified to address this. The equation here can be derived from the gradient wind equation by setting the Coriolis term equal to zero and allowing a strong local pressure gradient force to be balanced by centrifugal force at an arbitrary latitude ϕ.

Page 4, line 7; what is machine epsilon? Are you referring to machine precision?

They are quite similar but also technically distinct.

Machine precision is effectively the accuracy of the basic arithmetic operations.

Machine epsilon is the discrete distance between (for example) 1 and the next 'resolved' floating point number.

In this case, we use machine epsilon because we define convergence as the time when the 'distance' between iteration n and $n + 1$ is less than can be 'resolved' by the minimum gap between two floating point numbers.

Page 5, line 3; Why is the warm bubble hydrostatically balanced? The introduction and abstract highlight the importance of non-hydrostatic processes for testing global models, so this doesn't make sense. The vertical accelerations should be fairly small for this warm bubble relative to the supercell that is simulated with very large vertical velocities. The balance adjustment is a physical process that should be tested in the models.

The bubble is hydrostatically balanced in order to result in a smoothly-evolving solution at test case onset when the flow has not developed strong non-hydrostatic characteristics. Technically, there is no requirement that the bubble be balanced; however a less carefully-designed perturbation will result in gravity waves associated with flow adjustment in the first few timesteps and/or a less realistic supercell evolution. However, the long-term behavior of the solution is largely insensitive to whether or not the field is rebalanced.

Page 5, line 13; Should say "second-order diffusion operator with a constant sub-grid scale viscosity (value) is applied...". Are you using the same values in the horizontal and vertical dimensions? Is that appropriate for the grid cell aspect ratio?

Changed to 'as resolution is increased for a given model, a second-order diffusion operator with a constant viscosity (value) is applied to all momentum equations.'

The same value is used in the horizontal and vertical directions (unless specified in the relevant passage in the text). Since this diffusion is added to mimic turbulent dissipation within supercells, the choice of the same value for all three dimensions should be reasonable and is consistent with Klemp et al. [2015].

Page 5, sentence starting with "ICON applied. . ."; This doesn't sound like a departure from the default described above. Also, is diffusion really applied to the mass continuity equation (rho) in ICON? This is not typically done.

ICON departed from the prescribed test since it did not apply any diffusion in the vertical. The formal definition of the test case includes diffusion on the three-dimensional momentum equations (vertical diffusion is only applies to the background state perturbation).

The developers of ICON confirmed that the inclusion of ρ in the list of variables where diffusion is applied was erroneous. This has been corrected and the developers greatly appreciate the reviewer noticing this oversight.

Section 3.2; This whole section is very subjective with no analysis to back up any of the claims. In order to comment on the reasons for the differences in the model runs, some analysis is needed.

We have added additional text to help buttress some of the hypotheses here. However, we should emphasize that the primary goal of this manuscript is not to do a formal deep dive into all model differences but rather define the test case and scope of solutions from modeling groups that participated in DCMIP2016. We have chosen to leave formal attribution studies to individual modeling centers (or groups of modeling centers) as model design choices are implemented when accounting for a host of considerations versus the outcome of a particular test case.

Page 7, line 11; How do you know that this is noise? This statement is subjective and no quantitative analysis is done to back up this claim.

This was poor verbiage on our part, we apologize. We have replaced 'noise' with 'small-scale structure' since we are not trying to argue these solutions necessarily contain numerical or physical 'noise;' spurious or otherwise.

Page 7, line 15; This is a very weak statement and should be removed. What is meant by "coupling between the dynamical core and subgrid parameterization" and why would this lead to these differences?

We have added text to clarify that this is a hypothesis and a target for future work. There has been previous work indicating that the coupling mechanisms between the dynamical core and subgrid physics parameterizations can drive sensitivity in solution output [Gross et al., 2018]. DCMIP2016 did not formally control for this, largely because it is very difficult to tightly specify a coupling framework that satisfies the multitude of different numerical schemes and software infrastructures used by various modeling centers. Rather, the majority of work investigating physics-dynamics coupling has been done within individual modeling frameworks, and is something a subset of participating DCMIP models will likely look into over the next few years.

Page 7, line 21; This is not a bulk, integrated measure of supercell intensity. How about showing the domain integrated kinetic energy or total energy?

For this test, integrated kinetic energy is a difficult value to evaluate, as the atmosphere is not at rest at initialization (and the storm dynamics' contribution to the kinetic energy budget is actually quite 'minimal' relative to the KE of the background environment – on the order of 1% or so). Most of the energetic deviation from the initial sheared state is due to the strong vertical motion associated with the intensifying supercell.

We chose maximum updraft velocity over the entire domain as a metric for two reasons. One, it is a tangible quantity that is commonly reported in both observational and modeling studies of supercells. Two, we apply the assumption that maximum updraft velocity is a first-order proxy for 'storm intensity' as measured by a more 'dynamical' quantity than rainfall.

However, we note that the language here was not precise and understand why it may cause confusion, so we have modified to read '... more storm-wide measures...'

References

Markus Gross, Sylvie Malardel, Christiane Jablonowski, and Nigel Wood. Bridging the (knowledge) gap between physics and dynamics. *Bulletin of the American Meteorological Society*, 97(1):137–142, Jan 2016. doi: 10.1175/bams-d-15-00103.1. URL `http://dx.doi.org/10.1175/BAMS-D-15-00103.1`.

Markus Gross, Hui Wan, Philip J. Rasch, Peter M. Caldwell, David L. Williamson, Daniel Klocke, Christiane Jablonowski, Diana R. Thatcher, Nigel Wood, Mike Cullen, and et al. Physics–dynamics coupling in weather, climate, and earth system models: Challenges and recent progress. *Monthly Weather Review*, 146(11):3505–3544, Nov 2018. doi: 10.1175/mwr-d-17-0345.1. URL `http://dx.doi.org/10.1175/MWR-D-17-0345.1`.

C. Jablonowski, C. M. Zarzycki, K. A. Reed, P. A. Ullrich, J. Kent, P. H. Lauritzen, and R. D. Nair. The Dynamical Core Model Intercomparison Project (DCMIP-2016): Results of the Moist Baroclinic Wave Test Case. *AGU Fall Meeting Abstracts*, December 2016.

Ming Ji and Frederick Toepfer. Dynamical core evaluation test report for NOAA's Next Generation Global Prediction System (NGGPS). Technical report, National Oceanic and Atmospheric Administration, 2016. URL `https://repository.library.noaa.gov/view/noaa/18653`.

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

Joseph B Klemp and Robert B Wilhelmson. The simulation of three-dimensional convective storm dynamics. *Journal of the Atmospheric Sciences*, 35(6):1070–1096, 1978.

Zhiming Kuang, Peter N. Blossey, and Christopher S. Bretherton. A new approach for 3D cloud-resolving simulations of large-scale atmospheric circulation. *Geophysical Research Letters*, 32(2), 2005.

Nils P. Wedi and Piotr K. Smolarkiewicz. A framework for testing global non-hydrostatic models. *Quarterly Journal of the Royal Meteorological Society*, 135(639):469–484, 2009.