Reply to reviewer 1

We would like to appreciate the valuable comments which help for improving the manuscript. In the revision, we clarify the metrics to quantitatively evaluate model performance and add some validation results concerning TC surface wind structure (Figures 5 and 6). Figures 1 and 2 are redrawn by using the result of Stage 2 only with 95% confidence levels as error bars rather than standard deviation. Careful analysis of simulated TC position revealed that some misdetection occurred for very weak TC cases. These cases are excluded from validation.
Point-to-point response are following.

That said, the goals of the manuscript are somewhat unclear. Is the manuscript purely describing the TYMIP-G7 project framework such that it can be referenced in future studies? Or do the authors seek to describe fundamental differences in model results and attribute them to different model configurations? The authors bounce back and forth a bit between the two and the analysis of TC forecasts (beyond mean statistics) is somewhat weak, particularly in the final quarter of the manuscript. The result that increasing resolution improves TC forecasts is not tremendously novel in the community. The analysis of the different structure of the forecasted TCs is intriguing (although require significantly more work in future manuscripts) but the authors only select a particular forecast cycle to perform analysis on, which seems tenuous (at best) given the spread in TC intensity forecasts discussed earlier in the manuscript.
My recommendation is for "major revisions." I think the authors would be well-served to tighten up the description of the simulations and model configurations, which would give a very clear citation for future papers using TYMIP-G7 data. In addition, while the authors do not need to explain why models perform differently (those are additional projects in and of themselves), it would be useful to have something more than a single forecast initialization analyzed, particularly for Figs. 6 and 8. My preference would be to present a mean structure over multiple forecast cycles and explain that these differences exist in these model configurations and require additional analysis in the future. I have elaborated on major and minor critiques below.

In this revision, we clearly state that the aim of the manuscript at the last of Section 1: describing the specification of TYMIP-G7 and a set of metrics, and showing results concerning the metrics. We deleted Figs.6 and 8 which showed simulated TC structure for a specific case, but added a composite of axisymmetric primary and secondary circulation at the mature phase of TC to discuss the difference in simulated structure of TCs. Yes, we need further works to make clear what causes difference in simulated TC structure. Thank you for the comment.

In addition, there are phrasings that are somewhat awkward and grammatically incorrect for an English journal. I have noted some below but it is not meant to be an exhaustive list. My recommendation would be for a native English speaker to proofread this manuscript thoroughly before resubmission.

We are so sorry for our English quality. We ordered an English editing service by Enago before resubmission.

There is very little that can be said about model differences based on single forecast experiments. While I am aware that this manuscript is not intended to explain all of the physical differences (of which there might be many, particularly within the subgrid parameterization suites), I am worried that there is little utility in Fig. 6 and 8. I would anticipate being able to find cases where, for example, TCs have more asymmetric structure in GSM (even with lower resolution) or look more like observations, due to the fact that there are many forecast cycles from which to pick from. The same goes for the depth and structure of the axisymmetric circulation. Picking single members from the envelope of Fig. 4 implies that you cannot adequately understand model differences because you aren't removing run-to-run variability. In Fig. 8, it's possible that the NICAM signal (TC with lower outflow jet and shallower inflow) is a physical signal (perhaps due to the NICAM setup itself) but it also may be that that particular forecast in NICAM had more vertical wind shear than the other model configurations. My preference here would be for there to be either multiple TCs explored or perhaps some sort of average across a number of forecast cycles (say, Fig. 8 could be the average of 20 different TCs at +96 hour lead time from 20 different forecast initializations).

Thank you for the comment. It should be very useful for further detailed analyses. In the revision, we deleted Figs.6 and 8 which showed simulated TC structure for a specific case, but added a composite of axisymmetric primary and secondary circulation at the mature phase of TC.

Why is only the second stage shown in Fig. 3 but both stages are included in Fig. 4? This is especially relevant since the authors state that "track errors in MSSG were larger than those of GSM" in Stage 1, which is the opposite of the Stage 2 results (Fig. 3). If the errors associated with precipitable water (Page 8, line 25) were severe enough to eliminate their usage in Fig. 3, why weren't they eliminated in Fig. 4? Also, *why are there error bars in Fig. 4 but not in Fig. 3? Error bars should be included in Fig. 3 to give a sense as to the spread around the mean. It is difficult to understand whether those differences in track are "significant" (in either a statistical sense or just by subjectively assessing the figure).*

We used the result of Stage 2 only for Figs. 3 and 4 and added error bars showing 95% confidence levels rather than one standard deviation.

The timing results are very underdeveloped. For example, what is "execution efficiency?" To be honest, I'm not sure if this adds a great deal to the manuscript. Timing studies seem most useful either a) when as many variables are constrained as possi-ble (i.e., same resolution, different physics, etc.) or b) operationally, when a wall clock time benchmark threshold is required. For example, here DFSM is much faster, so in an operational sense, a forecaster might say "why don't we just use DFSM?" However, a more rigorous timing analysis might want to demonstrate the strong and weak scaling properties of the model and what happens if different subgrid parameterizations are used. Furthermore, Table 5 currently investigates only one forecast cycle. Individual forecasts may have different timings (even with the same model) for a variety of reasons (different load on the computing cluster, how the communication is spread amongst nodes, failures/bottlenecks during I/O write to disks, etc.). My recommendation would be to just remove the table (since this is R2O) and spend a brief paragraph discussing mean timings (i.e., over multiple forecast cycles), but emphasizing that there are many, many different aspects of each model configuration that lead to the disparate timings.

Thank you for the comment. In the revision, we used computational resource for a 5-day forecast (node-hours) as a metrics to evaluate the timing of the model. The amount of resource is hardly variable among cases because computational nodes are occupied for a model experiment and disk I/O is performed from/to the work disk mounted on each computational node. We also discussed many aspects which affect timings.

The authors mention "errors" in Stage 1 forecasts multiple times during the manuscript but don't elaborate significantly. My preference would be for any changes/corrections between Stage 1 and Stage 2 that persist in the data to be noted clearly such that future analysis with TYMIP-G7 data can refer back to it (note, that if the authors corrected these issues and merely re-ran Stage 1 with the updated settings, there is no reason *to mention this as long as the "incorrect" Stage 1 data is overwritten).*

We decided to rerun the experiments in Stage 1 by MSSG using this year's budget but they have not completed yet. Because Stage 2 has enough samples to examine difference in TC track, intensity, and structure, we used Stage 2 throughout the revision. Since we believe describing causes of failure in MSSG would help some model developers, we remained the description. Thank you for your understandings.

- Page 2, Line 10: '... is to avoid that transform.' Please cite a reference.
- Page 10, Line 11: '... Skamarock (2004) stated that seven times...' this is dependent on the numerical scheme and not universal across all models. See Kent et al., 2014, JCP.

Deleted.

- Page 10: Line 28: Is there anything in this manuscript that evaluates rapid intensification (RI) critically? A figure such as Fig. 5 could be useful, but if Delta_SLP (change in surface pressure) is calculated, not absolute surface pressure. Otherwise, RI seems neglected, so I wouldn't include this as a main result.

We deleted some sentences concerning RI.

- Fig. 4., it appears the initialization is too weak across all models (~5 hPa), which could propagate through the intensity forecasts at long leads. This is particularly relevant for the DFSM model which is initialized too weak yet develops TCs that are generally too strong.

Thank you for the comment. Initial bias of TC central pressure is mentioned in the revision.

Grammar/Typos:
- Page 2, Line 27: '... form on annual average in the western North Pacific...' is awkward. Could be 'Since an average of 26 TCs (XXXX) form on average in the western North Pacific....'
- Page 2, Line 39: 'to' should be 'too'
- Page 3, Line 37: '... diurnal cyclone...' should be '... diurnal cycle...'
- Page 4, Line 5: '... most activate...' should be '... most active...'
- Page 9, Line 32-33: 'However, precipitation patterns...' should be 'However, the precipitation

patters...'

All comments concerning above grammar and spelling errors are corrected in the revision. Thank you for your careful review.