Reply to reviewer 2

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.

This paper provides a preliminary assessment of tropical cyclone quality obtained from three high-resolution models. Although the paper is well-written, understandable and provides interesting results, I am somewhat concerned about framing the paper as a model intercomparison. Namely, I would expect a paper that lays out an intercomparison effort would provide substantive details on how one can quantify success. This is particularly evident through section 5.3, where the assessment is purely qualitative and, by avoiding any negativity about particular modeling systems, fails to call out what seem to be deficiencies in the structure of the typhoon that arises in the MSSG and DFSM models. To address this concern, it is suggested that the authors explicitly callout the metrics (computational performance, track error, intensity error, structural error) that could be used to quantitatively assess model performance, along with how one could quantify success under these metrics. This listing is analogous to section 4.1, but with further quantification of success or error under each criteria. Further, a tabulated intercomparison of the compared models that shows successes / deficiencies would be helpful to the reader to more clearly see how they intercompare.

In this revision, we clarified metrics to quantitatively evaluate model performance (Section 4.1) and the evaluation results (Section 5). Thank you for the valuable comment.

Page 6, line 6: Why is the standalone MSSG-A used when a coupled oceanatmosphere version is available?

Because the other models are not coupled with full 3D ocean model, the standalone MSSG-A was used so far. As we discussed in Section 6, we would like to examine the ocean effect using AO coupled MSSG as well as AO coupled NICAM. Thank you for the understanding.

Page 7, line 11: Does the slab ocean model react appropriately to the passing TC by *generating a cold wake?*

The slab ocean model just calculates local heat budget between atmosphere and ocean slab. Therefore, no cooling due to vertical mixing or Ekman pumping, but cooling by shielding effect of short wave by clouds occurs (Page 7, lines 17-19).

Page 7, line 4-5: "split-explicit" and "horizontally explicit and vertically implicit" typically refer to different techniques. The former uses explicit sub-cycling to deal with some vertically propagating wave modes, whereas the latter uses an implicit solve for all vertical terms.

Thank you for the comment. This part has been corrected in the revision (Page 7, line 11).

Page 9, line 27: Is the appearance of the "wavy" structure of the tropical cyclone associated with Gibbs' oscillations that arise from the spectral-transform nature in the model?

Figure 6 was deleted in the revision.

Page 9, line 35-37: Some additional discussion should be provided here regarding correctness. It seems that NICAM is the only one that produces a structure that matches observations – is that a correct assessment? Since the paper advocates for this hindcast strategy for assessing model quality, it should be more explicit on how one can actually evaluate the models using a mechanism that is not purely qualitative.
Page 10, line 7: Again, what is "correct"? The focus on qualitative model differences in this paragraph gives no insight into actually evaluations of the model.
Page 10, line 17: Again, additional information is needed on model correctness.

In this revision, we clarified metrics to quantitatively evaluate model performance (Section 4.1) and the evaluation results (Section 5). Thank you for the valuable comment.

Page 10, line 11: This is actually highly dependent on the numerical methods employed in the dynamical core and associated diffusion scheme. For a model like DFSM one would expect a finest resolved dynamical mode closer to 4dx, whereas for NICAM, which uses a co-located finite volume method, one would expect a finest resolved mode closer to 12dx. See Ullrich (2014).

Thank you for the comment. This part is deleted in the revision.

Page 11, line 13: It is stated earlier that NICAM is used in coupled atmosphere-ocean mode, so this statement is not quite correct.

The slab ocean does not calculate any ocean dynamics like vertical mixing and advection (Page 7, lines 17-19). Here, we would like to state that coupling with a 3D full ocean model is needed to examine impact of ocean cooling by TC. The statement is modified in the revision (Page 11, line 25).