Review comments in BLACK

Author responses in Blue/Italics

We revised the manuscript to address all of these editor's suggestions. We also lightly edited the texts to remove typos and improve clarity. Our responses to the reviewer's comments are described below.

I recommend to address some minor technical/editorial corrections.

1. It is very difficult to distinguish three green marks in Fig. 7 (MPI-ESM, HadGEM2-ES and CESM1-BGC), even though the difference is one of the main topic in this paper. I recommend to use different kinds or sizes of markers for ease understanding.

   Thanks for the suggestion. We replotted Fig 7 using the 'black' color for the HadGEM2-ES. HadGEM2-ES can now be easily distinguished from two other two (MPI-ESM and CESM1-BGC).

2. The explanation of the significance and confidence level (in line 560-561), should be described in figure caption of Fig. 12.

   "All three linear regressions are statistically significant at a level >95%" is added to the caption of Fig. 12.

3. In Fig. S3, the R2 and RMSE values for HadGEM and CESM cannot be read. There are several definitions of R2 (for example, eight definitions are described in Kvalseth, 1985) and the values are different among the definitions except for the simple linear regression with intercept. Thus R2 is not suitable for comparing GCM results with observed data (note that the R2 values in Figs. 7 and 12 are fine because they are simple linear regression with intercept). I recommend to use the modelling efficiency (MEF) (Stow et al., 2009) instead of R2 for Fig. S3 to show model skill simulating the observed data.

   Thanks for the useful suggestion. We revised the figures to show all figure labels. We agree that the MEF may be better than the R2 in this regard. The modeling efficiency (MEF) is defined in Stow et al. (2009) as:

$$MEF = \frac{\sum_i^N (C_i^{obs} - \overline{C^{obs}})^2 - \sum_i^N (C_i^{obs} - C_i)^2}{\sum_i^N (C_i^{obs} - \overline{C^{obs}})^2}$$

where N is the number of observation, C_i^{obs} *and* C_i *are the i-th point of model and observations, respectively. Following this definition, we calculated the MEF and replaced the R2 with the MEF for each model in Fig. S3.*

Reference

Stow, C. A., J. Jolliff, D. J. McGillicuddy, S. C. Doney, J. I. Allen, M. A. M. Friedrichs, K. A. Rose, and P. Wallheadg (2009), Skill assessment for coupled biological/physical models of marine systems, Journal of Marine Systems, 76(1-2), 4-15, doi:10.1016/j.jmarsys.2008.03.011.