Dear Authors,

Thank you for having provided a careful set of revisions. Prior to accepting your manuscript for publication I still have two queries. The first is very minor and concerns the reply to Reviewer #1's comment on Line 106. It may be worth adding a sentence reflecting your reply in the text, as several readers may indeed wonder about this choice. The second concerns the statistical significance computation. I appreciate that you have now added a brief description of this, but I note that you do not mention whether/how you have taken into account the issue of multiple testing. In your geographical maps, you are conducting a very large number of statistical tests on spatially correlated data, which may lead to an overstatement of significance if no correction to the significance level is applied (see e.g. Wilks, 2016: https://journals.ametsoc.org/view/journals/bams/97/12/bams-d-15-00267.1.xml).

AuthorResponse: We thank the editro for providing additional feedback on our paper. We have now taken into account both points and revised the manuscript accordingly. For the first point: we have now included in the text justification related to the choice of using five-degree grid for regriding. The revised text reads as (lines 113:115):
"For skill evaluation, all model simulations were converted to a uniform five-degree grid in order to minimize effects from small-scale noise in the identification of large-scale predictable signals, recommended practice for evaluating the initialized climate predictions (e.g. Goddard et al., 2013)"

For the second point we have applied false detection rate to inorder to test for the multiple testing following recommendations from Wilks, 2016. Please see lines (133-134) which read as: "The statistical significance of the correlations and temperature differences is estimated by using a two tailed student's t-test. We apply the false detection rate (FDR) proceedure to test for multiple testing (Wilks, 2016) using $alpha_{CDR}$=0.1."