

## Comments by Referee #1:

(Referee comments are in black, while our responses are in blue)

A review of "Assessment of Ocean Bottom Pressure Variations in CMIP6 HighResMIP Simulations" by Liu, Schindelegger, Börger, Foth, and Gou

The authors compare ocean bottom pressure (OBP) variability from GRACE/GRACE-FO retrievals, bottom pressure recorder (BPR) observations, and CMIP6 HighResMIP Simulations for past and future periods. They identify regions where observed and simulated OBP time series agree or not, and they also highlight where simulated OBP variability changes substantially from the present to the future, offering interpretations in terms of physical oceanographic processes or observational considerations. The analysis largely focuses on periods from a couple months to a decade, and the authors perform their calculations both with and without the mean seasonal cycle removed.

This is a nice paper that presents a solid incremental advance in the science. As the authors explain, changes in OBP can arise from a variety of oceanographic and geodetic processes, so a study like this will be of interest to a wide community of geoscientists. What's more, the paper is well written, the methods and approaches are generally sound, and most scientific inferences are reasonable and justified. The paper should be published after minor revisions to address a few places where the reasoning could be clarified or the analyses could be expanded to make a stronger study. I thank the authors for making a satisfying study on a topic that's largely been overlooked in papers analyzing past and future climate model simulations.

Good luck,

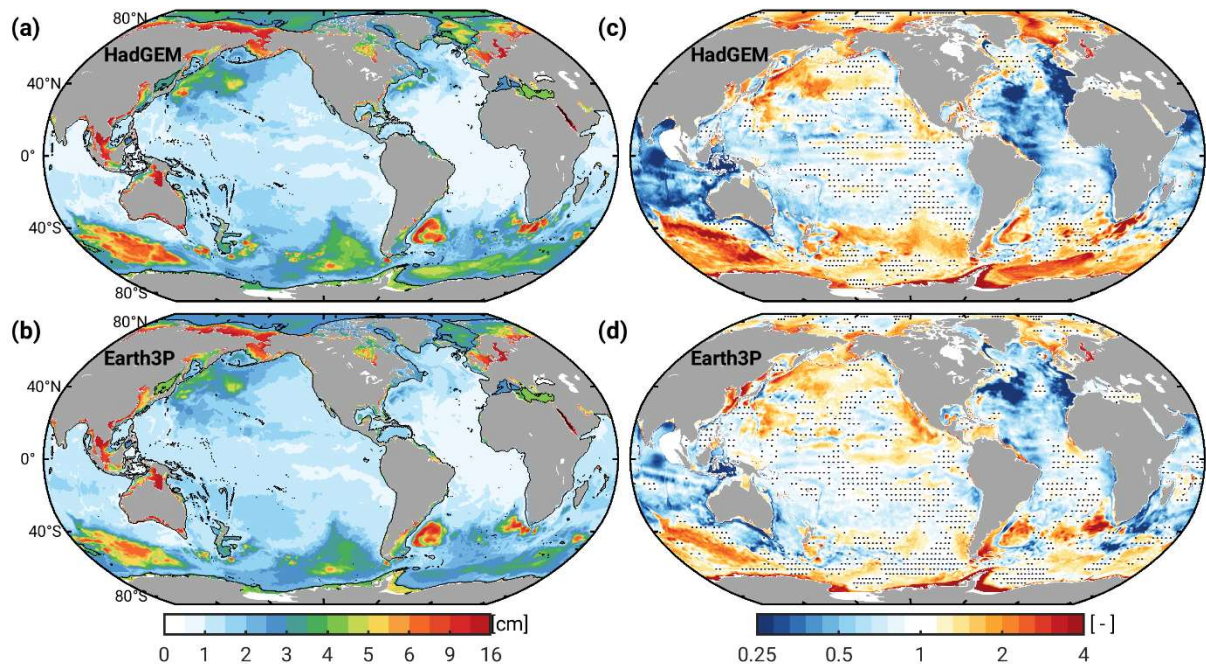
Chris Piecuch, Woods Hole

## General comments

Section 2.2 on calculation of OBP anomalies. Why don't the authors use the standard CMIP diagnostic output for OBP (pbo), which is readily available? I'd recommend them to use the proper model diagnostic output because, as the authors explain, right now they're making various assumptions in their calculation of OBP. The errors they're incurring from these assumptions are unclear. First, they're computing density from monthly temperature and salinity using the McDougall and Barker (2011) equation of state. While this sounds reasonable, it risks using an equation of state that's potentially distinct from what the various models use on line. It's also complicated by the nonlinearity of the equation of state (the monthly average of a density time series computed from instantaneous temperatures and salinities is not the same as the density computed from monthly averages of temperature and salinity time series). Second, the authors base their bathymetry on ETOPO1. Again, this sounds reasonable, but model bathymetry can be modified relative to something like ETOPO. Using standard model output alleviates these issues.

Thanks for the commentary. We analyzed two models at the start of the study: ECMWF-IFS 1/4° (see manuscript) and CESM1-3-HH 1/10° (which we dropped along the way). Neither

provides diagnostic ‘pbo’ output. We therefore set up our own calculation and applied it to other models without checking again for ‘pbo’. To follow your suggestion, we would either have to remove ECMWF-IFS from the paper or allow for a mix between model-true and inferred bottom pressures – neither approach is ideal. The good news is that the differences in RMS between ‘pbo’ and calculated bottom pressures are very small, visually barely distinguishable from each other. To assure readers, we have added the following figure to the Supplementary (Fig. S1):



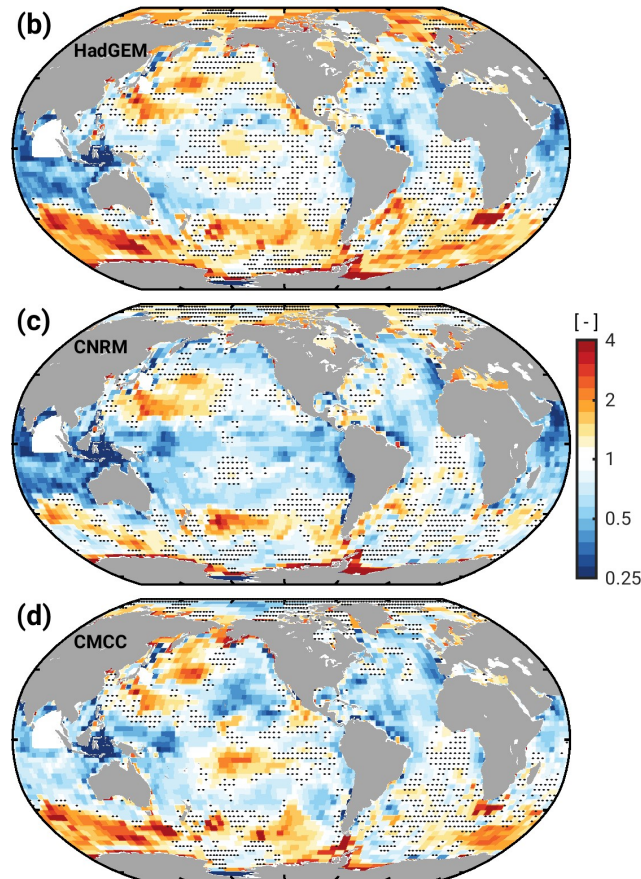
**Figure S1:** Plots on the left show the RMS (cm) of  $p_b$  time series from HadGEM and Earth3P, computed from the respective ‘pbo’ output over 1980–2014. Plots on the right show the variance ratio relative to the  $p_b$  variance from GRACE-DS. Stipples mark  $3^\circ \times 3^\circ$  cells where the RMS of GRACE-DS is within the 10<sup>th</sup>–90<sup>th</sup> percentile range of the model RMS for at least a third of the contained  $1/4^\circ$  grid points.

In eddy-active regions (e.g., Zapiola Gyre), the ‘pbo’ RMS is 0.2–0.4 mm smaller than the RMS of the calculated bottom pressures. Outside these regions, differences rarely exceed 0.1 mm. The availability of ‘pbo’ fields and the small differences seen in Fig. S1 are now mentioned at the end of Section 2.1 in the main text.

Section 2.3 on downscaling GRACE. I'd like the authors to discuss why they downscale GRACE/GRACE-FO data rather than coarsen CMIP6 model output. The downscaling is based on machine learning algorithms that incorporate eddy-permitting ocean circulation models. It's unclear what the associated uncertainties are. Are the downscaled datasets also based on the NEMO modeling framework? If so, then we have the potential of correlated errors between the downscaled GRACE/GRACE-FO products and the CMIP6 models. I'm not recommending the authors fundamentally change their approach. But I would like them to (1.) justify their decision to downscale rather than coarsen and (2.) discuss the associated uncertainties, biases, and other potential implications.

We have addressed this issue by repeating the model-data comparison (‘full’ time series) with two genuine GRACE datasets: (i) the CSR mascons that underlie the downscaled GRACE

product, GRACE-DS, and (ii) the coarse JPL mascons. For (ii), we averaged the CMIP6-HR fields to the exact same grid used by JPL (3° equal-area caps). Note that such coarsening is not feasible when comparing to the CSR mascons, because CSR employs a stochastic process model in the gravity field estimation that is difficult to mimic by simple spatial averages. The results of the new analyses are presented in Supplementary Figs. S2–S3 (variance ratio maps) and Supplementary Table S1–S2 and are briefly discussed at the end of Section 3.2 in the main text (‘Comparison with GRACE – broadband variability’). We include the middle part of Fig. S3 in this reply for illustration purposes:



**Figure S3:** Variance ratios of the full CMIP6-HR  $p_b$  time series from HadGEM, CNRM, and CMCC, with JPL mascons chosen as reference data and the model  $p_b$  fields averaged to the coarse JPL grids. Stipples mark 3° x 3° cells where the RMS of the observations is within the 10<sup>th</sup>–90<sup>th</sup> percentile range of the model RMS.

In short, the validation of the coarsened model output against JPL confirms the previously noted overestimation (underestimation) of modeled OBP variability in energetic (quiet) deep-ocean regions. Global median variance ratios,  $\bar{R}$ , for the energetic case are essentially identical to estimates of  $\bar{R}$  with GRACE-DS as reference. For the quiet deep-ocean case, JPL suggest  $\bar{R}$  values that are somewhat ( $\sim 0.1$ ) higher than with GRACE-DS. We attribute this difference to uncertainties in the various gravity field solutions (CSR, JPL) in the Atlantic; it is not something the downscaling introduces. On continental shelves, the original CSR mascons are an alternative benchmark, and here again, we observe only small changes (0.01 to max. 0.15) in the variance ratio statistics compared to the evaluation against GRACE-DS.

To prepare for these analyses in the text and motivate them in light of the possible limitations of GRACE-DS, we have expanded Section 2.3 ('GRACE data'). We have also modified Fig. 1 to show, along with the GRACE-DS RMS, the RMS of the CSR parent solution. This visual comparison allows one to (i) grasp the OBP variability added by the downscaling to the GRACE data, and (ii) appreciate the similarity between the two datasets at typical GRACE wavelengths. The justification for using GRACE-DS as main validation data simply reads "... we focus on GRACE-DS, as its signal content and characteristics are most commensurate with the CMIP6-HR models across scales and from deep to shallow regions."

Equation 2 (the variance ratio). The authors define this as the modeled variance divided by the observed variance. When they show results on a linear vertical scale, this definition will tend to visually overemphasize values  $>1$  and de-emphasize values  $<1$  (i.e., the former will span a greater color range than the latter). Therefore, I'd suggest, whenever they're showing R values, the authors use more of logarithmic color scale or vertical axis. That way, values  $>1$  and  $<1$  would visually communicate comparable emphasis.

Very good idea; we have implemented this in all figures showing variance ratios.

In all the figures, the authors compare root mean square (RMS) variability from the models and the observations. Typically, modeled RMS values are computed for a fixed time period (e.g., 1980-2014). However, Figure 5 shows the very interesting and important result that, even under a control simulation with presumably stationary statistics, there can be large apparent changes in RMS amplitudes (see left column). Because of this stochastic variability, it's unclear whether any of the R values in the preceding figures are significant or not. Therefore, I'd like the authors to perform a more comprehensive error analysis. Rather than computing modeled RMS values over a single period (e.g., 1980-2014), I'd suggest the authors instead compute RMS values for overlapping but separate periods to approximate a distribution of RMS values that would better quantify uncertainty and permit them to test whether simulated OBP variability really is distinct from what we're seeing in the observations. Thanks a lot for this helpful suggestion. We have added a paragraph to Section 2.2 ('Calculation of ocean bottom pressure anomalies') that introduces the non-stationary character of the model RMS to the reader, along with our approach to reflect this in the presentation. We closely followed your idea of creating an RMS distribution from rolling (in our case 15-year) windows within the longer time period, 1980–2014. In figures that show variance ratios model-versus-GRACE, we stipple grid points where the GRACE RMS is inside the 10<sup>th</sup>–90<sup>th</sup> percentile range of the model RMS. See Figs. S1 and S3 above for examples. In addition, in plots and for all quantifications, we now use the median of the distribution to infer the RMS of a particular model at each grid point (previously: RMS over the full 1980–2014 span). This led to minor changes of the global median variance ratio estimates.

Note that these basic error considerations based on RMS distribution only pertain to the assessment OBP output from historical simulations (Sections 3.2–3.5 in the revised manuscript). For the analysis of projected OBP changes in the models, we had already implemented proper statistical tests (see Fig. 5).



Starting on line 228, the authors note that models show stronger OBP variability on shelves compared to observations. They may mention that this could arise partly if the models aren't frictional enough or are too shallow (e.g., you expect barotropic ocean response to scale with the inverse of both friction coefficient and ocean depth).

Thank you for the insights; we have added this argument to the new Section 3.4 'Intermediate discussion', which we have created to accommodate suggestions by Referee #2 concerning the interpretive framing.

In section 3.5, the authors argue that increased future OBP variability could be related to changes in zonally averaged wind speeds (Figure 6). To me, this is an apples to oranges comparison. Assuming a linear barotropic adjustment such that OBP responds to winds on the same timescale, the more relevant comparison to make here would be to compare RMS (not mean) zonal wind speeds between the two periods. That is, the authors should quantify if the winds will grow more variable in the future (not just if they grow stronger overall).

As for the increase in OBP variance in eddy-rich regions of the South Atlantic, it is meaningful to analyze changes in the overall wind speeds, as their intensification and poleward shift appear to be the root cause for increased Agulhas leakage (e.g., Biastoch et al. 2009, Beech et al. 2022 and references therein). In the Arctic (our other regions where make a serious connection to surface winds), one should look at the wind speed RMS, that is correct. We have done the necessary analysis and created a preliminary plot of changes in wind speed variance for inclusion in the Supplementary. In all coupled model simulations that show a significant increase in OBP variance across the Arctic Ocean and Nordic Seas, we find enhanced future wind speed RMS in some of the forcing regions for Arctic OBP variability (see Fukumori et al. 2015).

#### Line edits:

Line 34: Suggest to delete "mainly because they are arbitrary in time"  
Phrase deleted.

Line 50: Suggest to delete "by us"  
Done.

Line 51: Suggest to change "Given the monthly sampling of the data and the fact that model drift precludes the study of trends" to "Given the monthly sampling of the data and the fact that models drift, we are precluded from studying trends"

We now write "Given the monthly sampling of the data and possibility of spurious model drifts, we are precluded from studying trends and exclusively analyze pb signals at intraseasonal to interannual periods."

Line 68: Suggest to change "Ad hoc short names" to "Abbreviations"  
Done.

Equation 1: Change " $\int_0^{\eta} \rho_0 g dz$ " to " $\int_0^{\eta} \rho g dz$ " and change the second equal sign to an approximation symbol (since the authors make the very reasonable approximation that density is constant over the vertical distance between 0 and sea level)

Done, thank you.

Line 97: Change "sketchy" to "uncertain"

Done.

Lines 173: Specify *\*planetary\** potential vorticity

Done.

Line 184: Suggest to change "baroclinic instability" simply to "instability" to be more general

Done, thanks.

Line 208: Suggest to change "are therefore likely" to "may be"

Done.

Line 262: The concept of geostrophic modes (resonances) is fairly specific (and esoteric; see Greenspan 1968). Suggest to change "geostrophic modes" to simply "variability".

Done, thanks.

Line 274: Suggest to change "from this behavior" to "to this behavior"

Done.

Line 275: pb should be italicized

Done.