Comments by Referee #2:

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

This paper evaluates ocean bottom pressure (OBP) variability in high-resolution climate model simulations submitted to CMIP6 under the HighResMIP protocol. The authors compare model-derived OBP variance fields at 1/4° resolution with observation-based estimates from downscaled GRACE satellite data and in situ bottom pressure recorders (BPRs). Their results suggest the models overestimate variance in some regions (notably on continental shelves and the Southern Ocean abyssal plains) while underestimating it in more quiescent deep-ocean areas relative to observations. Future scenario analysis indicates a projected increase in OBP variance in high-latitude and eddy-active regions, which the authors link to enhanced wind forcing and intensified Agulhas leakage, suggesting interesting implications for satellite gravimetry and climate change detection.

With new gravity missions planned by ESA and NASA, the paper is timely and addresses a gap in our understanding of how high-resolution climate models represent ocean bottom pressure variability, and will be of interest to the oceanographic and geodetic communities. Overall, the manuscript is well written, the analysis is well-grounded and the figures are of good quality.

A significant concern, however, arises from the choice of reference data used to assess model performance. The GRACE-DS product, while innovative, is downscaled using ocean reanalysis outputs from GLORYS and ORAS5, both of which are based on the NEMO ocean model. Since NEMO also forms the ocean component of all five CMIP6-HR models assessed here, the evaluation may suffer from circularity: structural biases present in NEMO-based models could propagate into both the GRACE-DS product and the CMIP6 simulations. This undermines the independence of the benchmark and makes it difficult to unambiguously attribute over- or underestimation of OBP variance to model error rather than artefacts of the downscaling process. Moreover, the over or underestimations may be greater than given by this analysis. A deeper discussion of this limitation, or a sensitivity test using alternative GRACE products and/or reanalyses, would help clarify the robustness of the findings. This may be beyond the scope of the present work, but potential limitations should be more fully acknowledged. And, in light of these issues, it would be more appropriate to describe the results in the more neutral terms of relative amplitudes rather than the loaded terms 'overestimation' and 'underestimation'. This shift in language would reduce the implication that one dataset is definitively correct and better reflect the comparative nature of the analysis.
We understand the skepticism and agree that potential limitations of GRACE-DS and resultant implications for the analysis were not properly addressed in the original manuscript. The concern largely overlaps with the second point raised by Reviewer #1, so we write a very similar response: As a sensitivity test, we have repeated the model-data comparison ('full' time series) with two genuine GRACE datasets: (i) the CSR mascons that underlie the downscaled GRACE product, and (ii) the coarse JPL mascons. For (ii), we averaged the CMIP6-HR fields to the 3° equal-area caps used by JPL, which allows for a cleaner comparison at typical GRACE wavelengths. The results of these 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').

In short, the validation of the coarsened model output against JPL confirms the previously noted characteristics of modeled OBP variability in energetic and 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 (~0.1) higher than with GRACE-DS. We attribute this increase mainly to differences 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. Overall, these cross-checks support our initial conclusions based on GRACE-DS and allow us to attribute high and low relative amplitudes in CMIP6-HR to model errors, and not to artefacts from the downscaling.

Related points & modifications:
- Section 2.3 ('GRACE data') acknowledges possible limitations of GRACE-DS and prepares for the use of the CSR and JPL mascons in the supplementary analyses.
- Figure 1 has been updated to show the bottom pressure RMS from both GRACE-DS and the CSR parent solution.
- Given the successful checks for robustness, we still use the terms 'overestimation' and 'underestimation' in our presentation. However, we agree that these terms are somewhat loaded and have adopted alternative formulations in about half of the cases.

A broader concern is the paper's tendency to offer speculative explanations for inter-model differences and model–observation mismatches without direct supporting evidence. While many of the proposed mechanisms—such as topographic smoothing, wind stress misrepresentation, blocked ice shelf cavities, or changes in eddy activity—are plausible, they are presented more as assertions than as tested hypotheses. For example, differences in bottom pressure variance are attributed to bathymetric constraints or wind forcing without showing comparative diagnostics of wind fields, bathymetric detail, or eddy characteristics across models. One exception is for the Arctic and South Atlantic where there is a rather superficial attempt to relate long-term changes of OBP variances to the wind field, yet this is not convincing or properly developed. Similarly, interpretations of future OBP variance increases invoke processes like Ekman pumping, Rossby waves, or stratification changes, but these remain unexamined. This interpretive style, while common in model evaluation studies, risks overreaching and may give a false sense of causal understanding.

(comment continued)

I suggest removing the speculative content currently embedded in the results section, collating and synthesising it in a separate discussion section. This would clarify the distinction between empirical findings and interpretive hypotheses and improve the scientific rigour of the paper. While this restructuring might leave the results section relatively brief, it creates an opportunity to deepen the quantitative analysis — for example, by more systematically evaluating inter-model spread, introducing uncertainty estimates for variance ratios, or providing more regional or temporal breakdowns of the comparisons. A clearer separation between results and interpretation would also help readers better assess the robustness of the conclusions.

Thanks for the critical comment. We agree that our approach of presenting results – starting each paragraph with quantifications and closing with interpretation – was not ideal. We have dealt with this issue as follows:

- We have removed several arguments that were stretching it too far. The invoked processes were coastally trapped waves (Section 3.3 'Non-seasonal pb variability'), suppression of mesoscale signals on the continental slope (previously Section 3.4 'Comparison with in situ observations'), and Ekman pumping, baroclinic Rossby waves, and changes in stratification (previously Section 3.5 'Projected changes in pb variability').
- Other thoughts on possible model limitations that lead to differences with the satellite data (e.g., model resolution, shallow bathymetry, momentum transfer and dissipation schemes, blocking of ice shelf cavities) are now presented collectively in an intermediate discussion (new Section 3.4). After that, we move on to the comparison with BPRs, where one more interpretation (on topographic smoothing) is offered. This is the only manuscript structure that works considering the scope of the paper and the need for a natural flow of the presentation.
- In formulating our interpretations, we now mostly use the subjunctive. This reduces the risk of giving the reader a false sense of causal understanding.

We also would like to note that even if the new intermediate discussion is disregarded, the quantitative results section in the manuscript is not short. Region-specific information is already contained in the figures, while uncertainty information for model-data differences and validation against original GRACE data have been added during this revision. At the time of this writing, we are working on an improved version of Fig. 6 to bolster our arguments about the connection of changes in OBP variance to changes in surface wind speed. Apart from these new elements, we do not intend to add more results or drop individual sections to deep certain aspects of the quantitative analysis.

Given the novelty and relevance of the topic, and the sound core methodology, I believe the paper has the potential to make a valuable contribution. However, the concerns outlined above—particularly regarding the choice of reference dataset and the interpretive framing—should be addressed through major revision.

Minor comments

Line 56: "valorizations" -> "evaluations" or similar.
Done.

Line 96: "sketchy" -> "uncertain".
Done.

Line 142: reflecting *net* atmospheric pressure variations over the ocean
Done, thanks.

Line 216: “considerable” -> “somewhat”
Done.

Line 214: “This type of…” - this is an important caveat that needs stating upfront.
We have moved the caveat regarding mesoscale OBP signals in the GRACE-DS product to the data description section (Section 2.3).

Line 250: This is a weak justification.
There is no compelling reason to favor any particular model from the available pool (ECMWF, HadGEM, CMCC). We have changed our justification to “… a somewhat arbitrary choice but still a counterpoint to the other two models.”

Line 258: Must be the case.
Agreed, but we have removed/reformulated this passage.

Line 273: Why not discuss in more detail?
GRACE-DS has been previously validated against BPRs in Gou et al. (2025). Moreover, because the manuscript is about OBP signal content in the climate models, such discussion would rather be a digression.

Line 274: “signal levels” -> “amplitudes”.
We see nothing wrong with the word “signal levels” here and in other places of Section 3.

Line 275: pb italicised.
Done.

Figure 5: This caption is rather confusing.
We modified some of the wording to improve clarity.

Line 347: Why should decelerated winds lead to a reduction in Agulhas transport and eddy activity?
That was a fallacious argument, indeed. The cited papers (e.g., van Sebille et al. 2009) suggest that the key element is a repositioning of the latitude of zero wind stress curl due to the poleward shift of westerly winds over the Southern Ocean. We have reformulated the passage to something less specific: “… global warming is expected to strengthen Southern Hemisphere westerlies and progressively shift them to the south (Deng et al. 2022, see also Fig. 6). This change in the large-scale wind forcing will in turn reduce volume transport and eddy kinetic energy in the Agulhas Current, …”