Response to reviewer #1:

In general my comments from the previous round of reviews have been addressed. There are still some minor grammar errors but, as stated last time, these should be captured at the proofing stage.

I do have two final queries on the manuscript:

Line 215: again, I would ask the authors to check that they do mean to refer to Armitage and Davidson (2014) and not Armitage et al (2016) here

Figure 7: I found it difficult to see the two different shades of green contours. While in general the lighter contours, for 95% confidence, fall within a large contour of 68% confidence, it can be hard in some places to tell if this is the case or if it is indicating a region that falls below 68% confidence. I would suggest making the light contour even lighter (or the 68% confidence a darker green) to help with this, or using an alternative colour such as magenta for the 95% confidence contour.

The two comments of reviewer #1 has been adressed and corrected.

Response to reviewer #2:

General comment:

This paper is not necessarily about GRACE. It is just that GRACE has limitations and that all other studies use GRACE for closing the sea level budget, while this paper choose a different approach that allows comparisons into the pre-GRACE era. The main concern about GRACE is that there is not one GRACE-solution, but GRACE-based mass estimates are constrained with a-priori knowledge which is interpreted differently across GRACE solutions. It is beyond the scope of this study to make an inter-GRACE comparison. Nevertheless, we have added two GRACE-products (JPL and GSFC RL06 mascons) to parts of the analysis.

BPR data are not useful for investigating secular mass-changes due to drift and changed geographic location (Proshutinsky et al, 2018 (JGR)). Trend estimates are very sensitive to small variations, so even though you could fit the monthly variation from GRACE to the BPR-record (like the 2010-paper of P-F), this would not guarantee that the BPR-trend is correct. Our results shows that GRACE doesn't add much information to understanding long-term sea level change in the Arctic and a theoretical approach (sea level fingerprints + IB) might as well be used (for long-term estimates). We have moderated the language accordingly at some places and don't claim that GRACE is worse than the approach used here. We just provide an approach to close the Arctic Sea Level Budget prior to 2003.

In my second review, I stated my dissatisfaction with the presentation of an immature manuscript and recommended rejection. I now review this third version, and I still do not find it to be publishable. Nevertheless, it is now at least in a reviewable form, so I will recommend "major revision" and state my reasons below.

The Abstract remains a list of things the authors have done; there is no statement of aim(s). Extension of record length is useful but not very interesting by itself. The authors have a lot

to say about the presumed deficiencies of GRACE products (fair enough) - so what, then, is their GRACE-related aim?

They generate new, separate records for changes in SSH due to steric changes and to mass distribution changes. The mass distribution changes rely fundamentally on tide gauge records. Tide gauge records are (by their nature) coastal. A good fit to coastal sea level is not evidence of good fit over the whole Arctic domain, however; at best it might be considered relevant to the shallow shelf seas. The reason for interest in GRACE in the Arctic is that it provides data on the deep, central part of the ocean. The authors cannot (as they do at the end of the manuscript) claim that their fields can validate GRACE when their fields are not validated away from the coast. There does exist a third resource apart from GRACE and their fields (independent, therefore) in the form of long-term bottom-pressure recorders in the Beaufort Gyre. They should use these (there may be more, maybe one in the North Pole observatory, I'm not sure about that) to check on their product in full ocean depth away from the coast.

There are two other good papers by Peralta-Ferriz (GRL 2010 and 2011) that have a lot of insightful material in this regard. Imagine a mode of SL variability that is like a drum resonance - pinned around the edge and moving up and down in the middle. Tide gauges do not see it. P-F and GRACE do. This is relevant to the Ekman spin-up / spin-down mechanism (Proshutinsky Phil Trans 2015, originally Proshutinsky & Johnson JGR 1997).

So what they need to do to make this manuscript acceptable is follow up on their introductory assertions about problems with GRACE near land and actually quantify the improvements they claim for their method by first comparing their fields with BPRs as an independent check, and then comparing the GRACE mass fields with theirs. GRACE products are readily available, so show the differences between their fields and GRACE. Is it true that they can demonstrate improvement near the coast? By how much? Is it significant? What about over the deep ocean where they tide gauge validation is not relevant? If their product is relatively unconstrained over the deep ocean, is GRACE actually better there?

Answer these questions satisfactorily and they may have an interesting, useful and properly validated study.

One detail: "However, the significant change in the Beaufort Sea coincides with the transition from Envisat to CryoSat-2 and a inter-satellite bias in DTU/TUM Altimetry can not be excluded". Armitage (JGR 2016) treats Envisat / Cryosat crossover in the supplementary material, so their statement is false.

Yes – but this claim is about the DTU/TUM sea level product and not Armitage (2016). Also, even though the mean fields (not directly crossovers) are subtracted for the overlap period, it doesn’t necessarily mean that it removes the satellite bias when the overlap-period is happening in a time of rapid change (as in the Beaufort Sea in 2010).