Review of "A Bayesian approach towards daily pan-Arctic sea ice freeboard estimates from combined CryoSat-2 and Sentinel-3 satellite observations"
By Gregory et al.

General comments

In "A Bayesian approach towards daily pan-Arctic sea ice freeboard estimates from combined CryoSat-2 and Sentinel-3 satellite observations" the authors investigate the use of Bayesian inference to produce daily gridded pan-Arctic radar freeboard estimates. Gaussian Process Regression (GPR) is used to model spatio-temporal covariances between observations made by three ESA's satellite altimetry missions (CryoSat-2, Sentinel-3A, and Sentinel-3B) and to make pan-Arctic predictions of radar freeboard, with uncertainty estimates, on a given day.
This is a novel, interesting and relevant investigation as it attempts, for the first time, to estimate freeboard with a daily temporal resolution based solely on satellite altimetry data. The improved temporal resolution of pan-Arctic freeboard could contribute to our ability to understand physical processes driving sea ice thickness variability on sub-monthly time scales.
The study is generally well structured and the manuscript is clear and pleasant to read. I recommend this paper for publication, however, there are some points that should be addressed by the authors first.
Thank you for your kind words, and for taking the time to review our work! It is very much appreciated. Please see our comments below, which we hope address your concerns.

Specific comments

Data

- Why did you choose data between December 2018 and April 2019? By selecting e.g. the following season (2019/20), you could have included in the analysis the months of October and November and make your results representative for an entire Arctic winter season.
  The choice to perform our analysis for the 2018-2019 season was initially to compare with the final Operation Icebridge campaign in April 2019, however as we state in the manuscript, it was difficult to draw any conclusions based on so few data points. Note that we do plan to run this algorithm for future seasons and make the data publicly available in the near future.
- L85-90: Hamming-weighting and zero-padding are both applied to CS2 L0 processing (https://wiki.services.eoportal.org/tiki-download_wiki_attachment.php?attId=4431&page=Cryosat%20Documents&download=y). Please amend this statement and, if CS2 L0 data are processed using GPOD, please state the differences with the official Baseline-D version provided by ESA.
  We have not compared the GPOD-derived CryoSat-2 radar freeboard with the ESA L2 baseline D product, however in Lawrence et al. (2019) they applied the same L1B -> L2 processing to GPOD L1B and ESA L1B (baseline C) data and found a radar freeboard difference of ~6mm attributed to the fact that the GPOD L1B data does not contain the stack standard deviation (SSD) parameter which is used for filtering lead and floe waveforms in the ESA L1B -> L2 processing chain. As the authors remark in their paper, it was more important to ensure consistency between CS2 and S3 than consistency between our CS2 radar freeboard and the ESA L2 freeboard product. However we agree that it is important to comment on this and we will note this difference in the revised manuscript and say that our combined product is preliminary and awaits the availability of ESA L2 Sentinel-3 freeboard which is processed in a consistent way to CS2.

Method

- How do you treat observations from different satellites in the same grid cell acquired on the same day (i.e. co-located in both space and time)? Do you include these as separated inputs or do you feed them as a single averaged estimate to the GPR algorithm? This should be clarified in the manuscript.
  Observations which are co-located in space and time are treated as separate inputs. The GPR framework assumes that these observations are independent random samples drawn from the same distribution (i.e., the same posterior function we are trying to learn), yet have independent noise contents. We will make this clearer in the revised manuscript.

- As there is no general "Discussion" section, I add this comment here. Bayesian inference allows to estimate the optimal covariance function hyperparameters based entirely on data as the parameters maximising the log marginal likelihood function. Do you think that the tool you developed could be useful in investigating the spatial and temporal correlation length scales of freeboard measurements? Please add a short paragraph discussing this possibility.
  Indeed, for each grid cell we do retain the learned hyperparameters which maximise the log marginal likelihood function. This therefore allows us to construct spatial maps of each hyperparameter (including space and time correlation length scales - see Figure 1 below). We agree that this information might be useful to potential end-users of this product and so will include a discussion on this in the revised manuscript.

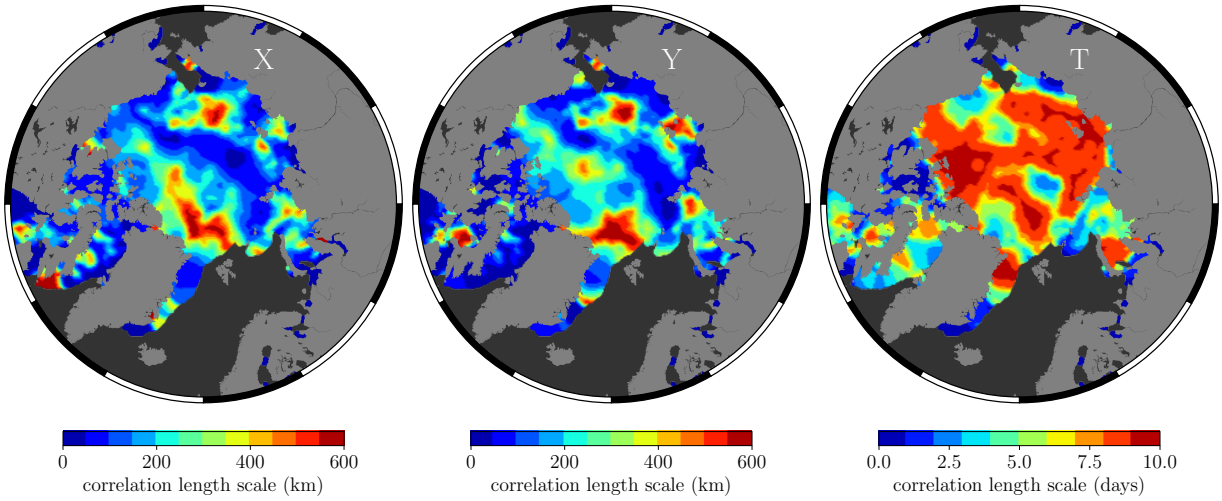

Figure 1 showing the zonal (X), meridional (Y) and temporal (T) freeboard correlation length scales which maximise the log marginal likelihood function, for each grid cell when generating predictions for the 1st of December 2018.

Validation

- How do you think a different grid resolution would affect your results in Section 4, e.g., by using a 25x25 km grid instead? Also, please repeat in the conclusions that the validation presented in Section 4 is based on a 50x50 km grid.
  In Figure 2 below, we show an example of the training error distribution for one day (1st of December 2018), where interpolations were run at 25, 50 and 100 km spatial resolution. Here we can see that increasing/decreasing the resolution does not result in a systematic

increase or decrease in the average error. We do however see that the spread in error is larger for finer resolutions, which is perhaps expected as the coarser resolution data will have averaged out much of the noise content within the data. On this note, we think that it is worth including some sensitivity tests as supplementary material for this manuscript, including tests where we vary the spatial grid resolution of the input data, and also vary the number of days in the training – see more below.

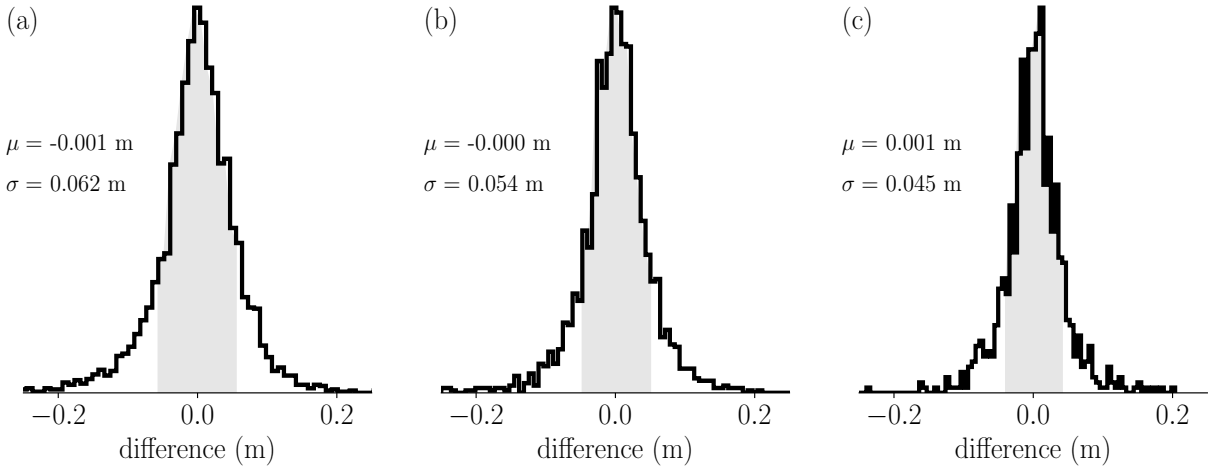

Figure 2 showing the error distributions between training data (CS2 and S3) and CS2S3 interpolated freeboards for the 1st of December 2018. Each distribution shows the error for interpolations performed at different spatial resolutions, with (a) 25x25 km, (b) 50x50 km, (c) 100x100 km.

- The results in Table 1 show a slight but systematically lower freeboard mean difference between CS2S3 and S3B compared with CS2S3-CS2 and CS2S3-S3A. While rounding might play a role in this comparison, do you have any idea why CS2S3 tends to best fit S3B data for every month of your analysis?

  Having since gone back and checked our calculations we have noticed a small error in the derivation of the mean and standard deviation statistics presented in Figures 4 and 5, and Tables 1 and 2. The revised statistics are given below for Tables 1 and 2:

Table 1 Date	μ CS2-CS3S3	σ CS2-CS3S3	μ S3A-CS3S3	σ S3A-CS3S3	μ S3B-CS3S3	σ S3B-CS3S3	RMSE CS2-CS3S3	RMSE S3A-CS3S3	RMSE S3B-CS3S3
201812	0.001	0.051	0.000	0.057	-0.001	0.057	0.051	0.057	0.057
201901	0.001	0.049	0.001	0.056	-0.002	0.055	0.049	0.056	0.055
201902	0.000	0.050	0.000	0.055	-0.001	0.055	0.050	0.055	0.055
201903	0.001	0.050	0.000	0.056	-0.001	0.057	0.050	0.056	0.057
201904	0.001	0.053	0.000	0.061	-0.001	0.061	0.053	0.061	0.061
all months	0.001	0.051	0.000	0.057	-0.001	0.057	0.051	0.057	0.057

Table 2 Date	μ S3A-CS3S3(-S3)	σ S3A-CS3S3(-S3)	μ S3B-CS3S3(-S3)	σ S3B-CS3S3(-S3)	μ S3A-CS3S3(-S3A)	σ S3A-CS3S3(-S3A)	μ S3B-CS3S3(-S3B)	σ S3B-CS3S3(-S3B)
201812	-0.002	0.073	-0.004	0.072	0.001	0.072	-0.002	0.072
201901	-0.001	0.071	-0.004	0.071	0.002	0.070	-0.003	0.070
201902	-0.002	0.072	-0.003	0.071	0.000	0.071	-0.002	0.070
201903	-0.003	0.074	-0.005	0.075	0.000	0.072	-0.004	0.0073
201904	-0.002	0.079	-0.005	0.076	0.001	0.076	-0.003	0.076
all months	-0.002	0.074	-0.004	0.073	0.001	0.072	-0.003	0.072

We now notice that CS2S3 freeboards are generally higher than S3B (given by the negative bias for both training and cross-validation comparisons, across all months). The model now appears to fit S3A better than S3B. Rounding does indeed play a role in these statistics, for example, if we increase the number of significant figures for the 'all months' cases $\mu_{CS2\text{-}CS2S3}$ and $\mu_{S3A\text{-}CS2S3}$ in Table 1, we see that $\mu_{CS2\text{-}CS2S3}$ = 0.00078 m and $\mu_{S3A\text{-}CS2S3}$ = 0.00024 m. Hence these round to 1 mm and 0 mm respectively. To address the question as to whether the difference in mean between any of the error distributions is significant (e.g., between $\mu_{CS2\text{-}CS2S3}$ and $\mu_{S3A\text{-}CS2S3}$), we can use a statistical Z-test. This can be computed through the following equation:

$$Z = \frac{\mu_{CS2-CS2S3} - \mu_{S3A-CS2S3}}{\sqrt{\frac{\sigma^2_{CS2-CS2S3}}{n_1} + \frac{\sigma^2_{S3A-CS2S3}}{n_2}}}$$

where n_1 and n_2 are the number of samples which make up the CS2-CS2S3 and S3A-CS2S3 error distributions respectively. The Z-test allows us to determine whether, based on the available samples from CS2-CS2S3 and S3A-CS2S3, the true means of the two error distributions are likely to be the same (i.e., the true zero-mean Gaussian noise distribution). Note that the Z-test assumes that samples are independent random variables – which is what assume the noise to be. In the example above we find that *Z* is equal to 2.38, which is equivalent to >99% significance. We therefore do not have evidence to reject the null hypothesis here, and can conclude that the two true means are highly likely to be the same.

- I understand the authors' choice of the cross-validation method, however, I think that both section 4.2 and the conclusions should clearly state that the given estimates of prediction error are based only on validation data from regions below 81.5.N and with a sea ice concentration larger than 75%, since these correspond to areas where the absolute uncertainty is usually the lowest (exception made for the Canadian Archipelago and the Fram Strait, as the authors nicely point out in Section 5). Regions above 81.5.N and with ice concentration between 15% and 75% (including the marginal ice zone) are systematically left out of the cross-validation since:
    - only S3 data are used as a validation
    - according to Lawrence et al. (2019a), diffuse waveforms within grid cells with ice concentration lower than 75% are discarded, which means that no freeboard estimates are available from any of the satellites on a given day where ice concentration falls below 75%.

  Thank you for raising this crucial point. We will make sure the manuscript reflects this in the revised version.
- I would have expected a more significant difference in performance when training the model with CS2 data only, given the lower spatio-temporal coverage when compared with a combined CS2/S3A/S3B training data set. According to your results, a GPR based on CS2 observations alone is able to predict radar freeboard at unobserved locations pretty well (with a 3-4% RMSE increase, from 5.9 to 6.1 cm, when compared to the multi-satellite solution). Do you think this is related to the relatively coarse (50x50 km) grid chosen for your cross-validation? I suggest to add a paragraph in your discussion elaborating on this matter and on the actual advantage of including S3 data in your model training compared with using only CS2 data. In the light of these results, it would also be interesting to discuss the possibility of using data from the three satellites while reducing the number of days used for model training.

With regards to the benefits of including Sentinel-3 data during the model training, we do see clear improvements in the derived freeboard estimates (see Figure 3 below). In particular, we notice how without S3 data, features such as the 'monkey tail' in the Beaufort sea are less well defined, and in some cases interpolation artefacts are present (particularly the CS2S3(-S3) case). Furthermore, we also importantly see reduced uncertainty in freeboard by the inclusion of all satellites in the training (see Figure 4 below).
With regards to reducing the number of days for model training, we generated sensitivity tests where we ran interpolations using 3, 5 and 9 days of observations during training (see Figure 5 below). Generally, we see that using only 3 days results in interpolation artefacts in some regions, which are significantly reduced (but not entirely eliminated) by increasing to 5 days. With 9 days of data, we see improved prediction performance and also, on average, reduced prediction uncertainty (see Figure 6 below). We will include a paragraph with some images on this in the revised manuscript.

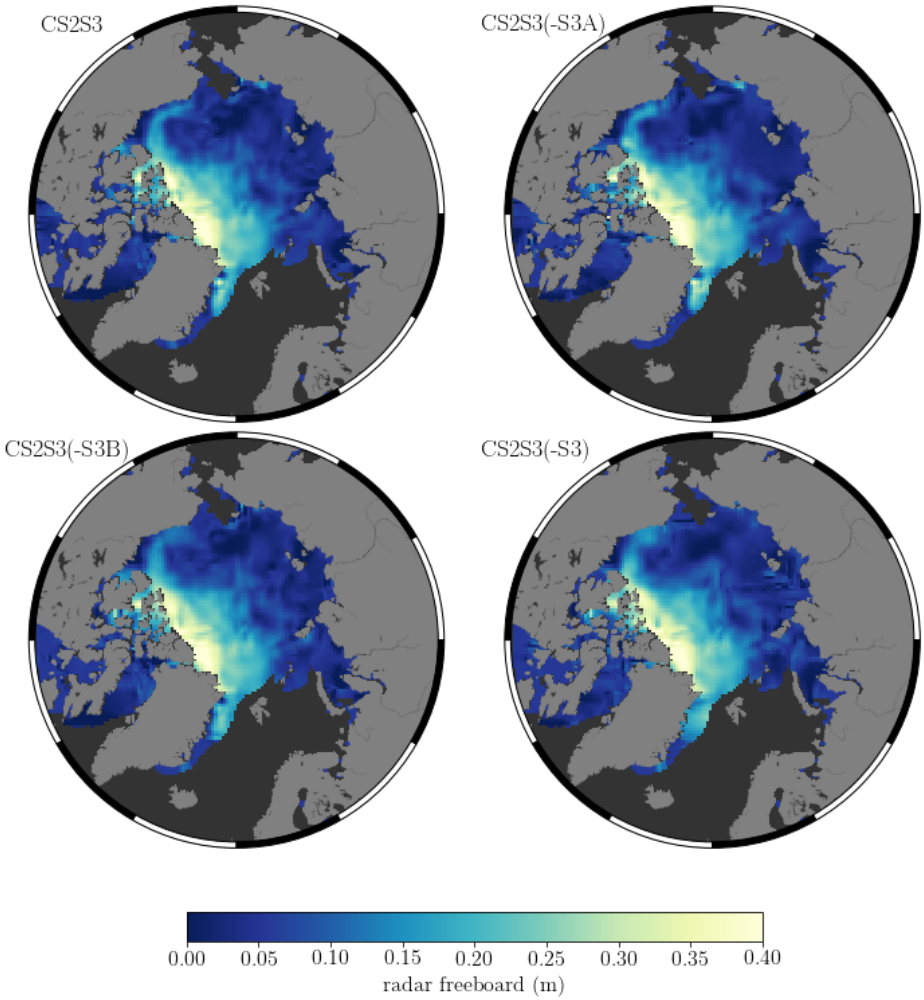

Figure 3 showing interpolated radar freeboards for the 1st of December 2018 at 50x50km grid resolution, and using 9 days of data during model training. Tests compare excluding different combinations of Sentinel-3 data from the training, as per the cross-validation tests of the manuscript.

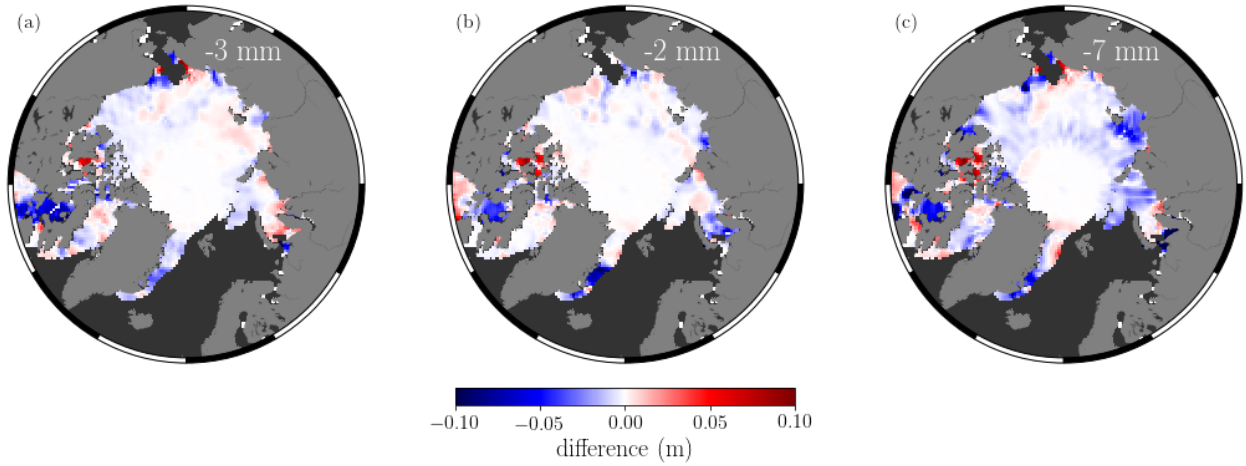

Figure 4 showing differences in prediction uncertainty relative to CS2S3 (trained with 9 days of data), for cross validation tests with (a) CS2S3(-S3A), (b) CS2S3(-S3B), (c) CS2S3(-S3). Blue colours represent lower uncertainty in the model trained using CS2 and S3. The average difference is given for each case (-3 mm, -2 mm, and -7 mm respectively).

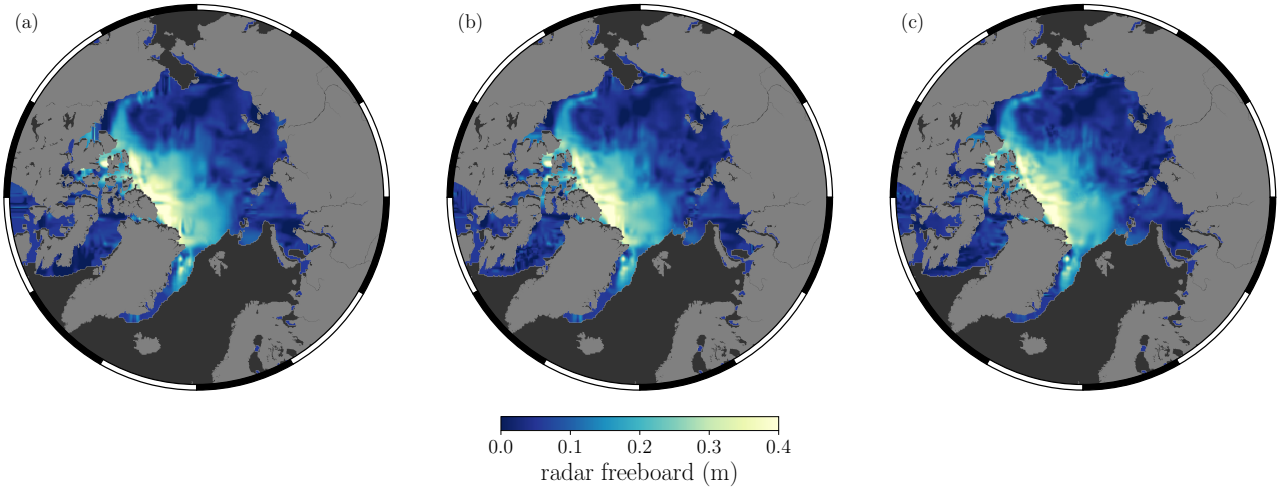

Figure 5 showing interpolated radar freeboards for the 1st of December 2018 at 25x25km resolution, where the interpolation was performed by using (a) 3 days, (b) 5 days, and (c) 9 days of training data.

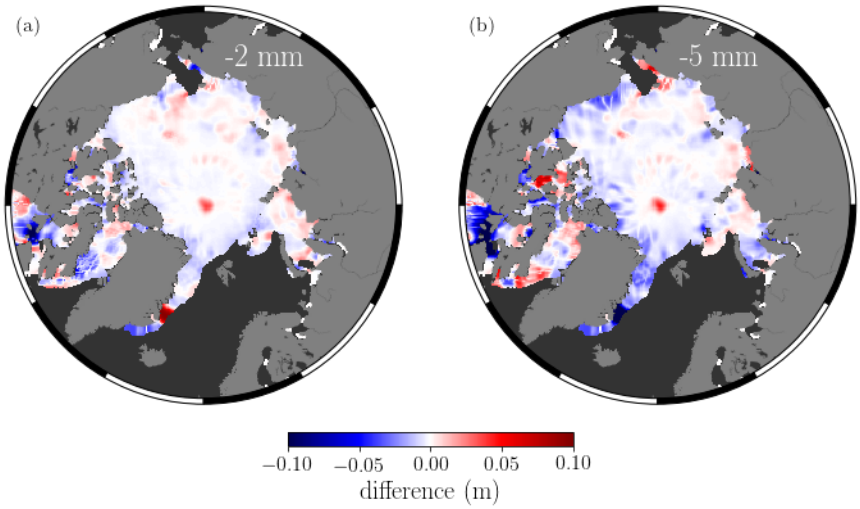

Figure 6 showing the difference in prediction uncertainty between the model trained with 9 days of data and (a) model trained with 5 days of data, (b) model trained with 3 days of data. The average difference is given for each case (-2 mm and -5 mm respectively). Note that uncertainty at the polar hole increases with more days of training data as no observations are ever recorded here.

Assessment of temporal variability

- This is a nice section highlighting daily variations of regional freeboard estimates and larger discrepancies between CS2S3 predictions and satellite data for sectors like the GIN and the CAA. I suggest to add a couple of statements about the 'Baffin & Hudson' sector. While the average CS2 and S3 freeboard over the entire period agree within 5 mm, they show differences of ~1 cm in December 2018 and March 2019. What do you think might be the reason for this more significant, with respect to other sectors, difference?
  Similar to the regions where we also see differences of ~1cm, e.g. GIN and CAA, we hypothesise that differences in the Baffin & Hudson sector are also a combination of lower latitudes, and therefore sparser sampling and higher uncertainties in interpolated sea level anomalies. We will incorporate this point into the revised manuscript.

- I would rephrase line 253 to reflect that the comparison of mean freeboard estimates over the entire observational period performed in this section is mainly a confirmation of your cross-validation results—the average value of a time series alone does not say a lot about temporal variability. I suggest something like: "… Generally, the mean of the CS2S3 time series lies within 3 mm of CS2 and S3, in line with the results of the cross-validation presented in section 4.2. However, ...".
  Thank you for the suggestion, we will amend this statement in the revised manuscript.

Technical corrections

We will amend all of the technical corrections below in the revised manuscript.

- L43: according to Lawrence et al. (2019a), the CS2 daily Arctic coverage is lower than 20% up to 82-83.N, not at all latitudes. Also, Tilling et al. (2016) shows Arctic coverage down to a minimum of two days, not one. Please amend this sentence to reflect the content of the cited publications
- L66/378: the DOI provided for Rasmussen and Williams (2006), a book, points to an article by Matthias Seeger with same title. Please correct the reference
- L104: if you want to be consistent with the platform/sensor notation used for the OSI SAF product, this line should perhaps read: "… from the Nimbus-7/SMMR, DMSP/SSM/I, and DMSP/SSMIS, which are ..." → (see https://nsidc.org/data/nsidc-0051 for reference)
- L108: you probably mean OSI-403-c? The 403-b product has been superseded and did not include AMSR-2 data
- L138: "For now..." → "For now, ..."
- L190: "corresponds" → "correspond"
- L291: add comma after "Greenland" → "… and the Greenland, Iceland and Norwegian Seas, ..."
- L301: I suggest not use "K" in the final statement → "… and the fact that the covariance structure can take any form, so long as the covariance matrix is symmetric, positive, and semi-definite, means ..."
- Figure 1: please state which day the sea ice concentration, type (FYI/MYI boundary) and radar freeboard refer to in the example
- Figure 3: please add the grid resolution (25x25 km) and the day which the radar freeboard estimates and uncertainty correspond to
- Figure 6: if the benchmark time series is not explained in the caption, please add a reference to the section 5

- Figure 7: please write the name of the sectors in full and provide the abbreviations, when used in the text and/or in Figure 6, in parentheses

  We would like to again thank the reviewer for their invested time in reviewing this work.