Reviews

We thank the Reviewers for the careful and constructive comments. The suggestions and corrections have greatly improved the quality of this manuscript.

Referee 2

The title says it all, really. I think that there is a lot of nice work here that should eventually be worth publishing, but it needs some thought, more work and revision before that is possible. My detailed comments follow, in order of occurrence.

2.1 p 1 L 8 delete "most" – a signal is either dominant or it is not

We changed the text, accordingly.

2.2 p 1 L 11 correlation coefficient r or r^2?

We changed the text, accordingly and added: **"Pearson linear correlation coefficient"**. The coefficient is referred to r.

2.3 p 1 LL 15-17 "deeper comprehension of the Arctic's DOT" – this sentence is empty because the manuscript is a technical identification of similarities and differences between measurements and model – it offers no new insights into the behaviour of the Arctic Ocean (not in itself necessarily a problem).

We fully agree with the reviewer, this was a formulation problem. This sentence was meant as an outlook in order to show a potential application of this study. We removed the sentence from the abstract.

2.4 p 1 Abstract: it is not clear to me that this manuscript contains properly formulated aims and objectives. It reads like a "look-see", and it does not need to; eg L 7 "to investigate similarities and discrepancies". If measurements and model agree, they are (probably) both right; if they disagree, you then want to identify which is "right" and which is "wrong", or even whether they may both be (differently) wrong. The authors should think more about how they frame the manuscript, therefore. See further comments on this below.

In order to clarify the aim of the present study we added a new sentence:
"The goal of the present paper is to identify to what extent pattern and variability of the northern Nordic Sea level derived from measurements and model respectively agree with each other."

2.5 p 1 L 20 "freshwater inflow" – false as expressed – altimetry plus geoid can tell you about steric (density) changes, it does not specifically tell you about salinity (ie fresh-water).

We agree. Altimetry can only recognize the geometric changes, We removed “freshwater inflow” from the text passage.

2.6 p 2 LL 15-16 potentially a very strong motivation for the study: you do not start by gridding, as others have tended to do. What are the benefits of this approach? It is hardly mentioned further but the manuscript really needs to draw out these benefits – assuming that they exist. If there are no actual benefits, then that too is worth knowing. I usually prefer not to suggest further work, but this is a case where it is necessary: please show the difference between a typical gridding approach to the altimetry and your finer-resolution approach. What matters and where does it matter?

The reviewer is right in recognising that one motivation for our study is to avoid to degrade the quality of the altimetry measurements by smoothing effects, i.e. by gridding the data. The final aim of our current line of research is that we do not want to fill the data gaps (that we cannot avoid when using along-track altimetry) by interpolation but by inclusion of additional data from a model. Before the combination can be done, a careful comparison of both data sets is necessary, and that is exactly what is the aim of this study.
In principle, it would be possible to include a gridded data set (gridded by ourselves or taken from an external source) in this comparison. However, since it is not the aim of our paper to assess different DOT products (in that case an additional validation by ground truth or in-situ data would be necessary), or even to evaluate the performance of FESOM and profiled DOT, we think that the inclusion of a gridded product could be included in a future work, but would not add significant progress to this particular study. The reason is that all existing gridded products show some limitations that hinder a comprehensive comparison with FESOM. For example, the two most obvious products are

- Absolute Dynamic Topography (ADT) grids, produced by CLS and CNES. Distributed by COPERNICUS MARINE ENVIRONMENT MONITORING SERVICE (CMEMS) (http://marine.copernicus.eu)
- Monthly DOT grids, produced by Centre for Polar Observation and Modeling (CPOM), Armitage et al., 2016.

(http://www.cpom.ucl.ac.uk/dynamic_topography/)
Both have a monthly resolution and a spatial resolution significantly coarser than FESOM (0.25° and 0.75°, respectively). Thus, small scale features are completely missing. Moreover, ADT are lacking any information on sea-ice regions. Hence, both solutions are not able to resolve smaller surface currents.

We added some more information on our motivation to the end of the introduction:
“In the present study, **along-track** high-frequency DOT estimates of ESA’s Envisat as well as water level outputs of FESOM are used for a direct comparison in order to analyze the spatio-temporal correspondence and discrepancies. **The overall motivation for this is the computation of a spatially homogeneous DOT without the need of gridding methods that smooth the altimetry spectral data content. Instead of such an interpolation, the unavoidable data gaps should be filled with model information by a combination of profiled altimetry data and gridded model data. A careful comparison of both data sets is a necessary prerequisite for such combination.** The **present** investigation aims at exploring capabilities for a combination and exploiting the advantages of both quantities. In particular, it is evaluated if the model outputs can bridge periods when altimetry fails (e.g. due to sea-ice coverage).”

Moreover, a new paragraph on the theoretical advantages of along-track data in comparison to gridded altimetry data has been included in the discussion section 4:
“Due to its measurement geometry, satellite altimetry has a high along-track resolution, but data are scattered in time and space. In addition, in polar regions, an irregular sampling due missing data caused by sea-ice coverage must be taken into account. This can significantly influence the estimation of annual sea level variability as tests with simulated data with different sampling revealed (see section 3.1). However, an interpolation of the data set (as it is done by the majority of other studies (e.g. Kwok and Morison (2015), Armitage et al. (2016), Farrell et al. (2012)) could be avoided in order to conserve more high-frequency observations and spectral content.”

Last but not least, also the conclusions (section 5) have been updated:
“Thus, it seems reasonable to exploit the advantages of both datasets by a combination of model and **along-track** observations. **This will enable the derivation of a homogeneous DOT, equally sampled in time and space without the need of smoothing the altimetry measurements by gridding procedures.** In such an approach,….”

2.7 p 2 L 19 delete "to" ("in spite of difficult")

We changed the text, accordingly.

2.8 p 2 L 25 delete comma ("conclude that")

We changed the text, accordingly.

2.9 Paragraph starting p 2 L 29 the justification for the use of FESOM is OK but I would like to see a line or two of context. What other models (if any) exist in this class (meaning spatial resolution, inclusion of ice and ocean physical processes, etc.), and how does it compare with them?

We added a description of other models focusing on Fram Strait ocean dynamics:

"Another sea ice—ocean model setup with comparable resolution focusing on the same region is based on a Regional Ocean Modeling System (ROMS), applying a grid size of 800 m around Svalbard (Hattermann et al. 2016). The model setup is regional, and nested into a 4 km pan-Arctic setup. In terms of eddy dynamics, the ROMS and FESOM setups compare very well (pers. comm., T. Hattermann). A slightly coarser model with up to 2 km resolution in the northern Nordic Seas was described by Kawasaki et al. 2015."

2.10 p 3 L 2 "eddy-resolving" – in most of the Arctic at most times of the year, but not everywhere and not always, see Nurser & Bacon (Ocean Science 2014). Near-zero wintertime shelf-sea density gradients can reduce the deformation radius below even 1 km.

We agree with the reviewer that the model is not always and everywhere in the 1 km domain eddy-resolving. In Figure 2 of Wekerle et al. 2017, the ratio of mesh resolution and Rossby radius is shown, based on climatological data. Particularly close to the Svalbard coast, it is not eddy-resolving. The ratio is just an indication that the model is eddy-resolving in most of the model domain and most of the time.
We added: **"in most of the study domain"**

2.11 p 3 L 16 "study area" – could use a more accurate description because you include the Lofoten Basin, discuss the Barents Sea, refer to part of the Arctic Ocean north of Fram Strait. The simplest solution is to replace "Greenland Sea" by "northern Nordic Seas".

We thank the Reviewer for this suggestion. We changed the title and several passages in the text.

2.12 p 3 L 16 reference to Figure 1. The figure is poor if you want it as a circulation sketch. Jan Mayen Current (south side of Greenland Sea recirculation), two branches of WSC, the baroclinic one runs further offshore

along Knipovich Ridge, what enters the Barents Sea, the polar current around Svalbard? There are plenty of such sketches around.

Figure 1 was updated to show all major currents. We also added a second plot displaying an averaged sea-ice concentration based on monthly NSIDC sea-ice concentration grids.

2.13 p 5 L 2 query as to meaning: "The model does not include...tidal changes". Do you mean that the model does not include tides? Please be explicit – this is important.

Yes, we confirm: FESOM does not include ocean tides. We modified "tidal changes" to **"ocean tide variations"**. We also changed "ignores" to **"does not include"** to emphasize the missing tidal variations. Moreover, the missing tides are now also stated in section 2.3.2 as a motivation for tide correction of altimetry data:

"One important correction is the ocean tide correction, since the FESOM model does not include ocean tides. In this study, we use EOT11a (Savcenko et al., 2012; Savcenko and Bosch, 2012) to correct for tidal effects. Even if EOT11a is a global ocean tide model it performs reasonable well in the Arctic Ocean (Stammer et al., 2014).

2.14 p 5 L 9 we are told that the model runs from 2000 to 2009. Your analysis start date is determined by Envisat, your end date by the model run. Please state this explicitly.

The FESOM model runs in the present configuration are available since 2000.
We started our analysis in 2003 due to inadequate first cycles of Envisat. Our study ends in 2009 because of missing model output for further years. We modified the text passage in section 2.3 and changed **2000** to **2003**. Furthermore we added:
"However, the first cycles of Envisat are affected by various instrumental issues and are not considered for the present study."

2.15 Section 2.3.1 starting p 5 it looks like you have a completely different approach to determining SSH in the presence of sea ice to the (by now well-established) Laxon method, but you say nothing about how or why it might be better. You really need to compare the two approaches, which I recommend you to attempt using the gridded product that I suggest you create above, and then comparing with publications that use the Laxon method. This might entail further analysis using EOFs, or calculating eddy kinetic energy.

We added additional information on the reason for choosing the unsupervised classification method as well as reference to some of the classical methods (Section 2.3.1):
"**Several classification methods have been developed within the last years, which are all based on the analysis of the returned satellite radar echo (e.g. Laxon 2004; Zakharova et. al., 2015; Zygmuntowska et. al., 2013). Most of them use thresholds on one or more parameters of the radar waveforms (e.g. maximum power or backscatter coefficient). In this study, an unsupervised classification approach is applied, which is independent from any training data. This method performed best in a recent study assessing the quality of different classification approaches with respect to very high resolution airborne imagery (Dettmering, et al, 2018).** Briefly summarized..."

As we also stated earlier in this rebuttal, creating a gridded product is not in the focus of this study. Surely, such a test would show differences originating from the altimetry classification but these will be mixed up with differences introduced by different altimeter data (missions and versions), different outlier detection, different correction models, and more. Thus, it will not allow for choosing the best classification method.

If the reviewer is referring not only to classification (section 2.3.1), but also to the SSH estimation by means of retracking (section 2.3.2), the comparison with the "Laxon method" was already done in Passaro et al. 2018, where Table 1 shows that the retracking method used in this paper (ALES+ on the table) outperforms both the ocean retracker (SGDR on the table) and the "Laxon method" (Laxon 1994a, Peacock and Laxon 2004) (SGDR-seaice) on the table. Passaro et al. 2018 performs a large comparison of these retrackers in part of the domain used in this study as well. The comparison in the Table 1 of the paper uses the Median Absolute Deviation between GOCO5s geoid heights and SSH data retracked with ALES+, SGDR-Ocean and SGDR-Seaice retracker in the test area.

2.16 Comment on Section 2.3.3 you use the "highly resolved...OGMOC" geoid. A conference abstract is not an adequate reference for this product.

We agree that a conference contribution is not the best reference. However, unfortunately, the geoid model is quite new and no peer-reviewed publication is available yet. The best reference we currently have is a recently submitted publication, which we added to the paper:

Th. Gruber, M. Willberg, Signal and Error Assessment of GOCE-based High Resolution Gravity Field Models, submitted to Journal of Geodetic Science, under review, 2018.

Moreover, it must be pointed out that OGMOC is nothing else than a combination of two well referenced geoid models: XGM2016 (Pail et al., 2018) for the lower harmonics and EIGEN6C4 (Förste et al., 2004) for the higher harmonics. For the spherical harmonic degrees 619 to 719 a combination of both models, using a weighting function, was performed. This information has now been added to the manuscript, in order to better describe the used geoid model. Please see section 2.3.3**:**

"Briefly summarized, OGMOC is a combination of XGM2016 (Pail et al., 2018) and the EIGEN6-C4 model (Förste et al., 2004). XGM2016 is used up to degree 619. Between 619 and 719, XGM2016 and EIGEN6-C4 are combined applying a weighting function. Higher harmonic degrees (>719) are retained unchanged to the EIGEN6-C4 model."

2.17 More importantly, and since I cannot tell how it created, I strongly doubt whether harmonics to generate product resolution below 10 km is at all meaningful. Satellite gravimetry can only "see" signals at around 100 km resolution; and if you are looking at Greenland shelf seas (as you are), the issue of "leakage" (terrestrial signal contaminating ocean signal) cannot be ignored. At present, it reads like you treat the geoid uncritically, as a "black box", which is not sufficient.

As already mentioned above, we modified the explanations about the model development and its individual components in order to clarify the structure of OGMOC (section 2.3.3). Please also note the last text lines of chapter 4 *Discussion*, which also describe the generation of OGMOC.

We agree with the Reviewer that satellite gravimetry can only observe spatial wavelengths above around 100 km. However, the applied geoid model includes beside satellite gravimetry observations, marine gravity information derived from various altimetry and in-situ observations (e.g. air, submarine campaigns etc.). This helps resolving shorter wavelengths.

Surely, leakage effects might degrade the accuracy in coastal areas (in addition to less accurate altimetry data due to land contamination of the signals). However, this problem exists for all global geoid models, and XGM2016 (and consequently OGMOC) shows improved results for near-coastal ocean regions with respect to other products (Pail et al., 2016). As long as no up-to-date regional geoid model is available, preferably based on locally supported functions like radial basis functions, this is the best model available today.

We are fully aware of these problems, but we would like to share with the reviewer our feeling that this awareness is shown in the manuscript: In section 4 *Discussion*, we discuss discrepancies between our two data sets resulting from the underlying geoid (last paragraph). Moreover, the conclusion clearly states the need for a better Arctic geoid.

2.18 Comment up to p 7. I have read to the end of Section 2 and there is nothing about tides, beyond a line in Table 1. Tidal corrections to altimetric SSH are

critical, and all tidal models have weaknesses in the Arctic and the northern Nordic Seas because the M2 cannot propgate freely north of the critical latitude – and S2 is aliased by sun-synchronous satellites. Use of EOT11a, however good it is globally, does not avoid this problem. Have you tried, as Armitage et al. (JGR 2016) did, comparing the model tides with tide gauges?

Of course, tidal corrections of altimetric SSH are not only critical but essential for the comparison with a non-tidal ocean model like FESOM. The aliasing the reviewer addresses is a principal problem of estimating tides empirically by single satellite data. EOT11a uses empirical estimates of multi-mission data and smoothly falls back to the numerical tide model FES2004 in areas not covered by satellite altimetry. All this is documented in the reference of EOT11a (Savcenko and Bosch, 2012).

The reviewer refers to the publication "Arctic sea surface height variability and change from satellite radar altimetry and GRACE 2003-2014" by Armitage et al. 2016. In this study, the authors perform a comparison against tide gauges using a single altimetry product. A validation of EOT11a against in-situ data, including the Arctic, has been already performed in Stammer et al., 2014 indicating that among other tide models, EOT11a performs rather good in the Arctic Ocean. In the meanwhile, there are newer models that came out in the most recent years and present some improvements in the Arctic, such as FES2014. But we still think that for the scope of this paper we can use EOT11a and the validation results of Stammer et al. 2014, which are much wider than any validation effort for tide models that we could do specifically in the context of this work. The following screenshot of table 6 (page 259) shows a snapshot of the validation results, published in "Accuracy assessment of global barotropic ocean tide models" by Stammer et al., 2014.

Table 6. RMS and RSS Differences (Both in cm) Between the Interpolated Tidal Signal and Common Tide Gauges for the Four Major Arctic Ocean Tidal Constituents[a]

	O_1	K_1	M_2	S_2	
No. of Stations	13	16	20	15	
Signal	4.81	10.35	30.46	11.45	RSS
GOT4.8	1.78	3.20	4.83	2.24	6.47
OSU12	1.77	3.17	4.82	2.22	6.43
DTU10	1.60	2.89	3.91	2.56	5.72
EOT11a	1.36	2.85	4.61	2.44	6.09
HAM12	1.25	3.00	4.11	7.11	8.83
FES12	1.52	3.30	4.66	3.36	6.80
TPXO8[b]	1.19	1.44	5.89	1.93	6.47

[a]The average amplitude (in cm) for each tidal constituents of the Arctic is listed. The values are given for the largest common set of tide gauges to five of the models.

[b]TPXO8 model assimilates a subset of the selected in situ measurements.

We added the additional information in section 2.3.2 as stated above (point 2.13) and attached following text:

"This study performs a validation by comparing different tide models to tide gauge data. For the Arctic Ocean, EOT11a shows RMS values between 1.4 cm and 4.6 cm for the four major constituents, and it is the second best of the seven models in the test."

2.19 p 9 L 14 you have identified a 3-day period artifact but you do not state what causes it; "irregular data sampling" is not an explanation.

The along-track sampled altimetry and simulated data show a 3-day period that cannot be explained physically. Due to this behavior we decided to add a comparison with the original meshed model data, where we couldn't identify this 3-day period (Figure 4b). For that reason we concluded that it must be related to the sampling or spatial distribution of the daily data.
We change the text passage to:
"This is an artifact **possibly caused by the data sampling. In order to prove this hypothesis, the frequency analysis is also performed for the full FESOM grid data.** Figure 4b shows …."

2.20 p 10 last line "bins of 7.5 km length" – state reason for choice.

This distance corresponds to the 1 Hz nominal ground track sampling of the ESA Envisat SGDR 2.1 dataset and simplifies our bin wise analyzes. Moreover, it reduces the computational efforts, we would have with another distance selection. We add "**1Hz**" and a some further information ("**… and reduces the high-frequency measurement noise**") to the text.

2.21 p 14 L 1 "These pattern originate from the altimetry DOT" – a factual statement without the implicit assumption would be "The patterns are seen in altimetric DOT but not in the model".

We changed the text, accordingly.

2.22 p 14 L 3 "insufficient sampling" – what does that mean, compared with my observations above about inherent weaknesses of geoids to do with resolution and leakage? What about tidal aliasing?

In fact, this sentence is misleading. In our opinion, this artifact is an error in the static geoid and most probably originating from biased altimetry observations in this region (due to sea-ice contamination or melt ponds). Since it is a single

structure in the open ocean, leakage can be excluded as an error source. Also tidal aliasing is unrealistic since it is a really small structure.
The sentence is now re-formulated. "…, these artifacts are due to **geoid errors caused by residual ocean signals at the polar latitudes…**"

2.23 p 16 L 6 here we are told that model lacks tides; this needs to be stated at the start.

This is already stated at page 5, line 6 in the description of the FESOM (section 2.2). Based on one of your previous comments, we now repeat this information in section 2.3.2.

2.24 p 16 L 6 and following, concerning barometric effects. SSH corrections include the inverse barometer (your table 1) – why is this insufficient?

At first, please note that there was a mistake in the 4th row of Table 1. We apply the Dynamic Atmosphere Correction (a combination of IB and high-frequent effects). We apologize for this confusion.

It is correct, that -in principle- correcting the altimetric-derived SSH by the IB effect should make the data set consistent to FESOM (the same holds for the tidal effect). However, model uncertainties of the corrections will show up directly in the differences and might influence the comparison. This was not precisely formulated in the manuscript. We changed the relevant paragraph:

Section 4:
"Furthermore, it **does not include** tidal ocean signal and barometric effects and is **lacking a steric correction to ensure the global conservation of mass. While the first two points are taken into account by correcting the altimetry observations, the latter point is currently not considered in the comparison. This should be acceptable since the impact on low frequency regional sea level patterns is small (Griffies and Greatbatch, 2012). However, it will contribute to the constant and long term differences visible in this study. In contrast, remaining differences in handling the atmospheric sea level pressure (i.e. caused by uncertainties of the used correction model) will show up in regional differences. They might be the reason for the observed temporal shifts of the maximum annual signal in the Greenland Basin. Even more important is the not sufficiently realistic consideration of freshwater inflow (e.g. by glacier runoff) by FESOM. This can cause phase shifts as well as reduced annual amplitudes. Furthermore the coarse resolution of atmospheric forcing is an additional reason for a smoothed sea level representation and an underestimation of annual amplitudes.**
For satellite altimetry…"

Moreover, we corrected the second to last paragraph of section 5:

"...FESOM should be **corrected for a global mean steric height change (Greatbatch, 1994) in order to ensure the conservation of mass and to make the observed altimetry heights directly comparable to the model heights. In addition, an improved handling of freshwater inflow is required to better account for mass changes due to glacier as well as river runoff.** However, even..."

2.25 p 17 para beginning L 15 you finally talk about tides, but I am not persuaded that you have investigated fully. AOTIM is a good regional (Arctic) tidal model (Padman & Erofeeva, GRL 2004), for example. But while satellite-based model currently suffer from the sun-synchonous problem (mentioned above), even good regional models contrained by tide gauges lack infomration away from the coast. What signature might identify, actually or at least hypothetically, unresolved tides in the altimetry?

At this point we have no simple answer on how to identify unresolved tidal signatures in altimetry. AOTIM certainly has been a good regional model outperforming TPXO6.2, both based on inverse modeling by the OTIS software. Meanwhile, version 1 of the high resolution atlas TPXO8 outperforms its precursor TPX07.2 in the Arctic by reducing e.g. the M2-RMS fit to some 244 tide gauge sites from 9.8 to 5.9 cm (http://volkov.oce.orst.edu/tides/tpxo8_atlas.html). But as the reviewer states, there is no way to verify the model performance away from the coast. Empirical models could help here, but suffer from aliasing that is only partly overcome by a multi-mission approach and crossover data.
To investigate which tide model is the best for the area of interest would be a study by its own. For more discussion about tides, we refer the reviewer to the section on p 17.

2.26 p 17 L 32 spell Greatbatch

We changed the sentence.

2.27 p 17 L 33 spell principal

We changed the text, accordingly.

So there is a lot of good work here, but I think that the authors need to do more to make this manuscript publishable. More context, more comparison with existing products and approaches, more thought about reasons for differences

between measurements and model, and not just leaning on the positive sides of the comparison. We learn new things where approaches disagree.