We thank the referees who have made excellent work in going through the details of our submitted MS and made very constructive remarks and corrections. Our detailed step-by-step responses to each of the Referee #2 comments or questions are given below.

We have revised the MS, with the following main points.

- The main points of the EOF reconstruction and the found modes were presented too briefly, relying mainly on the reference Elken et al. (2019). In the revised MS, additional important issues have been included in the compact form (hopefully not repeating the already published MS).
- Justification for the large-scale EOF DA method, in comparison with other well-known DA methods, has been refined.
- Data transformations between the fine and coarse grids have been more carefully presented.
- Unfortunately, the issue of observational errors has not been included in the initial MS. It is now included in the revised MS.
- Presentation of DA validation has been reformulated and discussed in more details.
- Possibilities of the method regarding operational forecast (with assimilating only the past data) have been discussed.

Suggested technical corrections have been included as well.

Anonymous Referee #2

Received and published: 15 July 2020

Comments and questions in bold
Response by the authors in normal
Line and Figure numbers taken from first submission

General comments:
The manuscript describes an unusual data assimilation (DA) method which employs Empirical Orthogonal Functions (EOFs) to correct only the large-scale patterns of an ocean model for the northeastern Baltic proper. A training dataset of five years of model data was used to calculate the EOF modes. Only sea surface temperature (SST) and sea surface salinity (SSS) are considered, and the method relies on observations from a time window of up to 30 days centred around the analysis time. The authors found that the DA method is feasible to use for assimilation of SST and SSS and that it is computationally efficient.

I think the authors have made an interesting investigation of the current setup using the so-called HBM ocean model and the proposed DA technique, and I recommend the paper be published after some corrections.

General remarks:
The manuscript is well structured but not always so easy to read. I recommend a language check by a native speaker, if possible.

We plan additional language check.

It is not clear whether the validation dataset was independent from the observations used in the DA process. On lines 239-240 it is stated that all observational data were used in the DA, but on lines 427 etc. it seems half of the gridded observations were reserved for validation. Please explain this more carefully.

The sentence on lines 239-240 was deleted as not necessary in this location.

In our method, DA depends only on the accuracy of observational gridded maps that were pre-calculated prior to the DA experiments. All the observations were included in the calculation. Experiments were made with options for reconstruction. Reducing the number of observation "boxes" by a factor of two gave nearly the same reconstruction results as the reconstruction with a full set of observational data. Pointwise comparison of SSS reconstruction over the full study period is presented in the Fig. X1 inserted below.

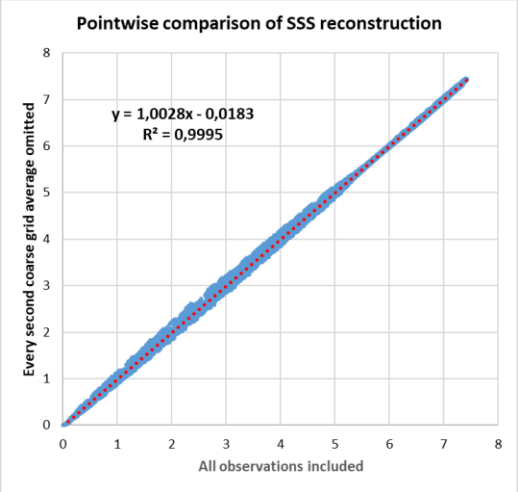


Figure X1. Scatter plot of all reconstructed SSS grid values over the study period: reconstruction with all observations included versus reconstruction with every coarse grid average omitted. Shown are the characteristics of linear regression.

The figure is not included in the MS, but the numerical estimates are given. The sentences on lines 426-431 are modified and replaced to:

"The experiments which took every second available observation "box" into account (this resulted in mean sampling interval along ship tracks about 20 km instead of 10 km) revealed that performing DA during the study period with reduced data set (6.5 k averaged observation data instead of 13 k) changed RMSD of SST by only 1% and of SSS by 2%, whereas the RMSD values were 0.05 °C for SST and 0.027 g kg^{-1} for SSS. It was evaluated over the full time span and domain using 182 k coarse grid cells; correlation between the data sets was higher than 0.999. We have also checked reconstruction results with FerryBox data only, excluding the data from shipborne monitoring stations. Compared with the full data set, largest (but still minor) differences with RMSD of SSS up to 0.03 g kg^{-1} were found in the Gulf of Riga and the eastern Gulf of Finland, where FB data were missing."

Also, even if every second gridded cell of observations were kept for validation, it is not clear to me that these observations are truly independent from the ones used in the DA, as they all originate from the same lines of FerryBox data. Is it possible to use data from certain ships for data assimilation, and use data from other ships for validation? Or is it possible to reserve the ICES data for validation, and just use the FerryBox data in the DA? Finally, the satellite-derived SST data in Figure 4 (d) is used to discuss the results in a qualitative way only; why not use it (together with similar satellite data) also for a statistical validation? In short, I would like to see a more careful validation using truly independent observations.

If the new observation points are separated by a distance of positive significant correlation, then the observational results are not truly independent. It is principally possible to use data from certain ships for data assimilation, and use data from other ships for validation, but the problem is that different ships cover different areas with different time intervals. For example, excluding the data from FinnMaid (Helsinki – Travemünde, Table 1) means that data from the Baltic Proper south from the Helsinki – Stockholm and Tallinn – Stockholm lines will be missing. It can be expected that different combinations of exclusion will give different results due to different geographical coverage. Sorting out such variations would require a large number of new time-consuming calculations, which is not reasonable for the first feasibility study of the method. We are updating both the computing facilities and the core operational forecast model, and plan longer DA experiments with more validation options in the near future.

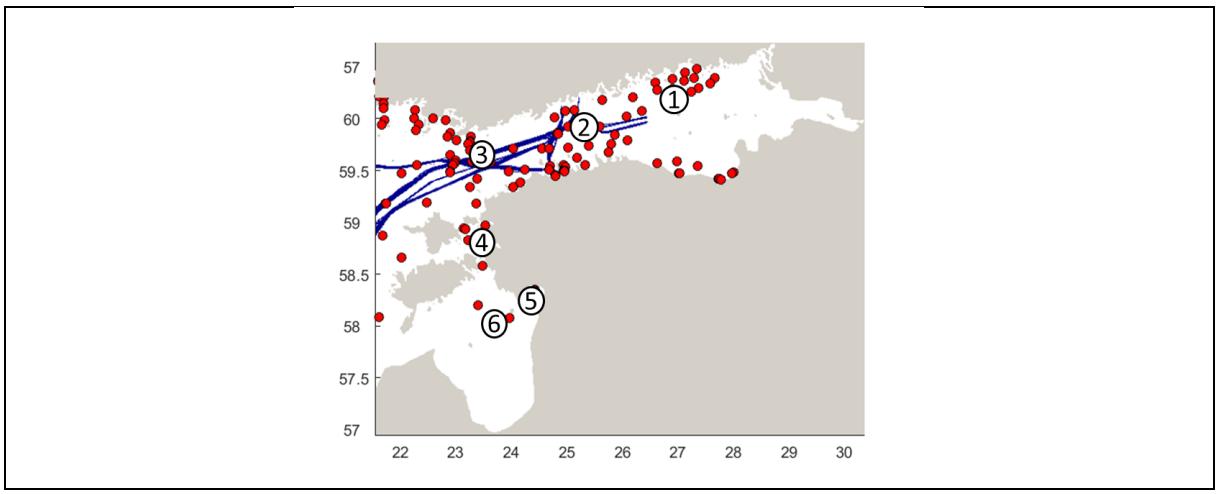
There were about 370 shipborne SST and SSS observations available, originating from about 80 spatially separated stations. This is a very small amount compared to the FerryBox data and therefore the shipborne statistics is not well comparable to that of the whole data set. We have done SSS reconstruction experiments as shown in Fig. X2. It was found that shipborne ICES data had only a minor effect on the results, since the large-scale variability with high spatial correlation dominates in the region.

We discussed earlier between ourselves about the possibility of comparison of SST DA with remote sensing results. We came to the opinion that this would bring too many details to our feasibility study, since it would also include non-trivial aspects of comparison of FerryBox data with different remote sensing products. This comparison can be done in a later stage

All the suggestions proposed are very valuable and we plan to perform such thorough validation studies at a later stage, when this DA system is going to be implemented in everyday forecast procedures.

We have added following text to the MS on line 431 before the last sentence of the paragraph:

"We have also checked reconstruction results with FerryBox data only, excluding the data from shipborne monitoring stations. Compared with the full data set, largest (but still minor) differences with RMSD of SSS up to 0.03 g kg^{-1} were found in the Gulf of Riga and the eastern Gulf of Finland, where FB data were missing."



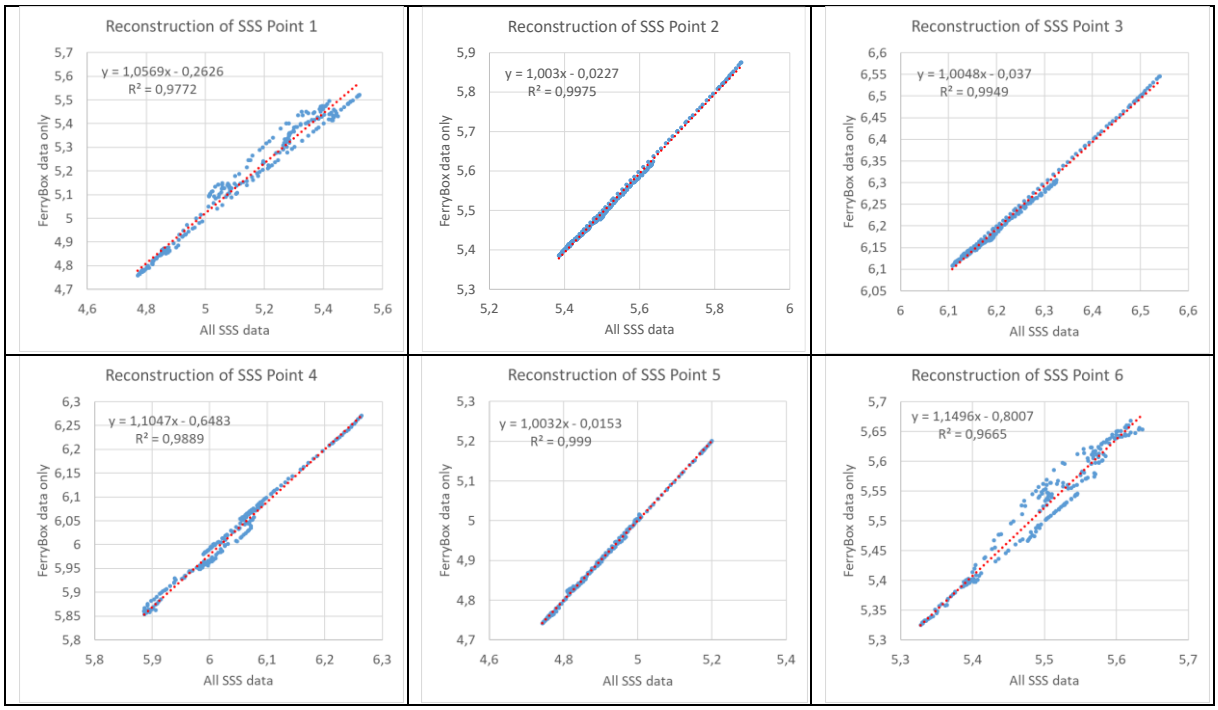


Figure X2. Scatter plots of reconstructed SSS time series at six locations shown on the map (on top of the panel). Shown are reconstruction based on all observational data versus reconstruction based on FerryBox data only.

The DA method relies on the use of EOF reconstructions of SST and SST. Please show some examples of the reconstructions that were used in the DA.

Examples of the reconstructions were added for 3 August 2015, to be compared with the maps in Figs 4 and 5 (former numbering).

It has been shown that the DA method works in a "reanalysis mode", in which observations "from the future" can be used in a time window up to 30 days wide, centred on the analysis date. I can see a problem when this method is used in forecast mode, where observations mainly from the past several days can be used. Is it enough to have a time window of, say, six hours? Please discuss this more.

Value of the time window depends on the spatio-temporal characteristics of the studied field and on the observational network. It is necessary that there are critical number of observations (say, 6 observations) available, in order to find observational EOF amplitudes from dominating modes. Remind that therefore we can detect only the large-scale patterns. With SST and SSS data, the amplitudes of dominating modes have temporal correlation scale generally more than 60 days, except for the SST “upwelling mode” which has about 15 days. We selected a centred time window of 30 days, although 10 days worked also well in most of the dates (some dates were dropped out because of too little data). If there are hourly time series available (like in recent years, there is data from buoy stations and gliders), it is possible to reduce the time window significantly, why not to try 6 hours. We did not consider in this study the sea level, but there is good hourly data available over all the coasts of the Baltic.

Time-dependent EOF reconstruction method enables the option to use only the past data as during the operational forecast mode. Time sequence of past observations is used to determine the rate of change of amplitudes, assuming that within the time window the amplitudes depend linearly on

time. There are good examples shown by Elken et al. (2019). However, extensive tests for using the past data only are not in the scope of the present feasibility study.

We have added following paragraph after line 487:
“We have tested the EOF-based DA in centred time window of 30 days, based mainly on available FerryBox data during the study period. As shown by reconstruction experiments by Elken et al. (2019), the time-dependent method can also work with backward observations as if it occurs during operational forecasts. When more observations become available, for example from new automated buoy stations, Argo floats and gliders, the time window can be shortened. Full covariance matrix estimated from the model results is the backbone of the EOF DA method. Prior and/or complementary to implementation of the method into operational practice, detailed covariance studies using results from multiple models could be useful, as well as additional reconstruction and DA studies using more data sources over longer periods.”

Detailed, technical comments:
(I omit page numbers as the line numbers are unique)
l.8-9: "...based on covariance estimates from long..." (too many "the")

Corrected.

l.10; "...on a regular grid."

Corrected.

l.35: "...do not presently include DA..." (word order)

Corrected.

l.66: "Baltic proper" (should not be spelled with capital P; see also other occurrences in the manuscript)

Both versions, “Baltic Proper” and “Baltic proper” are used in the scientific literature. Our historical preference of using “Baltic Proper” is partly reasoned by the HELCOM nomenclature of the sub-regions of the Baltic Sea, see https://helcom.fi/wp-content/uploads/2019/06/Implementation-of-the-BSAP-2018.pdf. We keep the term as it was written, “Baltic Proper”.

l.130: "Two sets of compressed (averaged) FerryBox..." (for clarity); also on line 133.

Corrected.

l.134: "...it was chosen not to..." (word order)

Corrected.

l.142: "...data was too irregular..."

Corrected.

l. 146: "...time fixed..." How do you mean? Are they time-independent? Or just interpolated to pre-determined, fixed times, e.g. 00 UTC?

The time-fixed approach uses EOF amplitudes that do not depend on time. Later, time-dependent amplitudes consider EOF amplitudes and their time derivatives within a selected time interval. For clarity, the sentence has been reformulated to:

"The basic option of EOF reconstruction uses at each DA time step time-fixed amplitudes, encountering the observations spanning over certain time (which can be longer than DA time step) that are transferred to the fixed times by some interpolation or filtering/averaging procedure."

l.170: "observation operator Hi..."

Corrected as suggested. Although, in most cases the operator takes the form of a matrix.

l.179-181: "In practice, ..." I don't quite understand this sentence; please rewrite...

This sentence has been deleted. The earlier sentence has been modified:

"Experiments with pseudo-observations (Elken et al., 2019) revealed that the values of $\hat{\mathbf{a}}_i$ of dominating L modes should match the limits derived from statistics of $\tilde{\mathbf{a}}_i$, whereas higher modes with outlying amplitudes should be neglected."

l.183: "...are not made at the same time." (simpler)

Corrected.

l.183: "...to cover a larger sea area..."

Corrected.

l.184: "...observation operator Hp..."

Corrected.

l.185: remove one instant of "that"

Corrected.

l.204: What happens if there is only one observation available?

With one observation available only, the amplitude of only the 1st EOF mode can be estimated, but most probably it will not fit to the statistical limits and have to be neglected. We have excluded the times when the number of observations was less than six (line 272).

l.215: "...without DA..." This puzzles me, as the analysis field depends on DA..?

The phrase has been rewritten:

"... calculated from the previous analysis field ψ^{a-1} ~~without DA~~ using only the model operator F without DA during this time step,"

l.224: "...can easily be included..." (word order)

Corrected.

l.271: "...frequently became unrealistic."

Corrected.

l.274: "The time windows for experiments (a) and (b) were selected to be 10..."

Corrected.

l.317: "...revealed a deep... ...and a shallower..."

Corrected.

l.344: "...strips of lower salinity..."

Corrected.

l.395: "...when DA had decreased the FR temperature..." FR = free run; so DA cannot affect the temperature... I think you mean "decreased the temperature in the SST01 and SST02 datasets"...?

Partially incorrect sentence was rewritten. The new sentence is:

"In the Gulf of Riga, SST increase dominated throughout the study period, but it was interrupted occasionally by basin-wide events when DA had decreased the ~~FR~~ temperature <u>compared to the results from FR</u>."

l.411: Is it the centred RMSD that is being used? Why not the usual RMSD? The centred RMSD is calculated after removing the bias; which do you mean?

We have used centred root-mean-square difference when comparing observations and model results. RMSD, giving also explanation "standard deviation of differences at all the observation points is denoted as centred RMSD". It means that the average difference between observations and model (bias) is not included in the centred RMSD. Many recent studies analyse the model results using Taylor (2001) diagram, which is based on the centred RMSD dependence on variances and correlation; our choice was made to have compatibility with such studies that consider bias and RMSD separately. Explanations of the RMSD acronym were checked throughout the MS and unified. In particular, the acronym RMSD was omitted when describing the use of least squares method.

l.418: "...so many observations."

The sentence has been reformulated:

"Areas with lower salinity in the eastern Gulf of Finland and in the Gulf of Riga ~~did not have any massive~~ <u>had only a small number of</u> observations."