Review of the manuscript

"Data assimilation of sea surface temperature and salinity using basin scale EOF reconstruction: a feasibility study in the NE Baltic Sea"

Authored by

M. Zujev, J. Elken and P. Lagemaa

General Comments

The paper addresses an important issue, which is the estimation of salinity and temperature for the Baltic Sea using a combination of model data and observations.

The method is based on a two-step approach, in which sparse observations are interpolated using an EOF technique and subsequently a relaxation method is applied for the assimilation into the model.

The method seems to have some potential for the assimilation of FerryBox data, where interpolation to 2D grids make sense, if longer time scales are considered.

There are a couple of concerns, which should be addressed:

- The method assumes in the interpolation step, that the covariance structure of the model is correct. This should be discussed more – in particular the limitations caused by this
- The previous point is related to a discussion of the main model error sources, which is missing as well
- Observation errors are not discussed at all – this needs to be justified and discussed
- We missed some discussion on the potential of the method to improve forecasts
- We would furthermore appreciate some discussion about the implications of the assimilation on the model dynamics (e.g., vertical density structure, in particular stability)
- It is not really clear, why the authors did not apply a more standard technique, with a more solid theoretical basis. A straightforward approach would be to use a low rank model error covariance matrix based on the presented EOF decomposition in the standard Kalman analysis equation. This would then also include observation errors and avoid the two steps required in the presented technique.

The presentation of the material should be improved. There are deficiencies, in particular with regard to putting the study into context of existing methods and motivating the selected approach. If "computational effort" is the main point, then this has to be quantified better.

There are quite a view grammar problems and a native speaker should proofread the text.

We recommend publication after major revisions.

Specific Comments

Abstract

We think it would be better to structure the abstract such, that more general information (what is done ?) comes first and specific results follow after that

Please explain acronym RMSD

I think it is more common to say "dominating EOF modes" instead of "gravest", but that should be checked by a native speaker

Introduction

Page 1, line 22: please reformulate "discrepancies of"

Page 2, line 34: replace "then" by "the"

Data and methods

Page 3, line 72 : "whichever" instead of "which"

Page 3, line 79: "… from the halocline …" please reformulate

Page 3, line 87: "better grid cells" instead of "points"

It would be good to learn more about the vertical discretization of the model, e.g., how thick is the surface layer.

Page 4, line 95, maybe better "grid resolution" instead of "grid step"

Fig. 2a: Please change color of FerryBox tracks – it cannot be distinguished from land.

Please add information on the water depth the FerryBox observations are usually taken.

Page 5: It was not clear, how you interpolate the FerryBox data to a 2D grid. Please explain in more detail.

Page 5, line 142 : Did you mean "… too irregular …" ?

Page 5, lines 146-150 : this paragraph is hardly understandable – please reformulate.

Page 6, line 160: "the" instead of "then"

Section 2.3

The entire section is unfortunately quite messy and confusing, although (as far as I understand) the method is quite basic. The authors have to explain all symbols and indices with much more care. Also, what is a vector or matrix (what size?) and what is a scalar ?

Page 6, line 185 : "that that"

The basic assumption, if you use EOFs for interpolations like you do, is that the covariance structure of the model is correct – this should be stated more explicitly and discussed a little.

You could have included observation errors in this interpolation exercise. I guess your assumption at the moment is, that the observations are 100% correct ? - please comment

Eq. 1: you assume that this matrix actually has an inverse – please comment. If the matrix is close to singular, you run into numerical problems as well.

Section 2.4

I guess eq. 1 is a continuous equation, which in its original form should be solved using the internal model time. I assume that you get eq. 4, if you replace the model time step by the assimilation time step – please explain more

I had problems to figure out how big the assimilation time step in the experiments actually is – please use consistent notation for cititical parameters (e.g. time steps) throughout the document.

Page 8: line 223 : "The DA method … is analogical …" I don't think this is true in this generality, because it seems you don't consider observation errors at all. – please comment. The resemblance with 4DVAR is remote, because there is no model dynamics included in the minimization of the cost function.

Page 8, line 239: "… artificial split …" I don't understand this sentence, because this "split" is a standard approach to validate assimilation techniques.

Page 7, line 217: " … since extensive use has been made …". This is I guess the critical point. The classical approach in an assimilation filter is to combine observations and the model state using covariance information on model errors at each analysis time step. In your approach there are no covariances of model errors. Instead, you use covariances of the background statistics for the interpolation. If you used a scaled version of the background covariance as a proxy for the model error covariance in a classical filter approach, you would probably end up with similar results, but with a more solid theoretical foundation. Anyway, as pointed out in the general comments, the method has to be put into the context of existing methods in a better way.

Section 3.1

Page 9, line 275: This is interesting; why don't you show the EOFS computed in your study ?

Page 13, line 408: The skill is often defined in relation to a reference run (e.g. the free run). In the case of the standard forecast skill, it is a dimensionless number – please check.

Page 16: "There are obvious extensions … layers …"
This is, where it gets interesting, because the vertical structure of different model variables (temperature, salinity, etc) is a particular challenge and your assumption about the correctness of model covariances may become a problem (e.g., if the mixed layer thickness in the model is not correct)

Figure 8: It would be interesting to see the absolute differences between observations and the assimilation run and the same for the free run (these differences should reflect both observation and model errors).