EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-539-RC2, 2022

Comment on egusphere-2022-539

Anonymous Referee #2

Referee comment on "Detecting anomalous sea-level states in North Sea tide gauge data using an autoassociative neural network" by Kathrin Wahle et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-539-RC2, 2022

Review of the paper "Detecting anomalous sea-level states in North Sea tide gauge data using autoassociative Neural Networks", by Kathrin Wahle, B. Emil V. Stanev, Joanna Staneva

The paper presents an interesting application case of Neural Networks (NN) for the assessments of non-linear ocean dynamics, which can be used to identify and classify interesting non-linear ocean configurations and to study the ocean model's ability to reproduce them in detail. It might be helpful for the development of adequate parameterizations and for process implementations in numerical ocean models. This I think is an interesting point of the paper. The method could maybe be used to analyze the model's ability to represent events with anomalous correlations across modelled sea level stations, but this is not in the focus in the paper, which is dealing with the identification and classification of measured events. However, hydrodynamic model evaluation is part of the paper, so the evaluation of the model reconstruction error can be seen to be in the scope of the paper as well.

I suggest the authors to go a bit more in depth with the results analysis in chapter 3, to compare the two storm events with each other and maybe to analyze the model's ability to reproduce the anomalous correlation events of sea level time series (even if this is not in the focus of the study). I also suggest to elaborate the method part of the paper, to provide more information about the NN method and trainings method, which would be interesting for readers unfamiliar with NN methods, and to provide an analysis of the weaknesses of the method. I think that this can gain the paper more interest.

My background is more in operational modelling than in artificial intelligence. My statements with regards to the use of machine learning thus represent the perspective of an ocean modeler who wants to apply the method and not so much the perspective of a developer, who is more interested in the method itself. My remarks on the applied method should be seen in this context.

I recommend major revision, because I think that most of the figures have to be re-done and a part of the manuscript needs to be re-written. The statement does not refer to the general scientific quality of the manuscript, which I consider to be good.

General comments:

For me as a reader with no background in machine learning algorithms and IA, it would be helpful to understand the rationale for choosing the method to apply neural networks and if other, more recent methods could have been used as well. Neural Networks use "supervised learning" to train the algorithm. Other, more recent methods use "deep learning" to train the algorithm. Could these be used as well? This could be of interest for readers from the operational modelling community, that are not necessarily experts in machine learning.

Neural network, testing and training method: Which method was used for training the algorithm? In the manuscript, the method is introduced and the results are presented, but the information on how the method was trained and has been applied is rather limited. Have k-fold cross-validation methods been applied? How was the time series sampled? It would also be interesting to know how long it takes to train the algorithm. The analysis bases on the comparison of the reconstructions from different AAN-Networks that were trained using different data sets. So, the time it takes to train one AANN is an important factor in the efficiency of the analysis.

Influence of observation errors, especially under storm conditions. There is usually a lot of jitter in the observational data set, especially at the time of the sea level peak. Is there a need for quality control of the input data set? Which influence has the time resolution of the training and testing data set. Is hourly time resolution enough, to study the non-linearities in the data set?

Could the authors provide some analysis of the weaknesses of the used method, maybe in the discussion section. Is it possible to run the method for any configuration of observation stations? What would happen if the array of observation stations was extended eastwards into the Danish Straits and the Baltic Sea, which are less dominated by tides? Would the method still work?

General points:

Use of English language: Please check the use of the english language and please re-write the text to improve the readability.

Figures: The quality of the figures needs to be improved, which might also involve the re-structuring of the figures. The resolution of the figures is fairly low. It is nearly impossible to read the text in the figures. Some figures have no title too, so that one has to guess what they are showing. Some of the sub-figures are not marked by letters.

Nomenclature: Please check if the naming is consistent throughout the paper. An example are "anomalous spatial correlations", which sometimes are also called "unusual spatial correlations".

My nomenclature throughout the review: In the text, "Line" with capital "L" refers to the line counter at the left side of the page, whereas "line" with small "l" refers to the line in the paragraph (line 6 in paragraph 2 of this chapter).

0.Abstract

Line 8-15, page 1: Abstract could be extended, to include all the topics covered by the study. The abstract in its current form serves mainly as a motivation. It is not describing the extend of the study, including the different application cases. I would also suggest to avoid specific, technical terminology or to re-phrase it, to not confuse readers with an operational background, which should be very interested in this paper. Terminology like "anomalous spatial correlations", "unusual spatial correlations", "lower dimensional subspace" should be introduced or re-phrased, to be better understandable.

1.Introduction

Line 27, page1: I would remove "largely".

Line 34, page2: The sentence about operational model data does not fit here between satellite altimetry and tide gauge data sets. It should be moved further down.

Line 37-39, page 2: Comment: When the model application for sea level observation data gap filing was mentioned, I thought first that this was what NEMO model results were used for in this study. It should maybe be mentioned at a later stage that this is not done.

Line 47, page 2: "Thermohaline forcing", is that coming from runoff of rivers?

Line 49, page 2: What is a “perfect model”? Is this a model without prediction error? I would reformulate the sentence using a different word.

Line 58 (and below), page 2: The term “anomalous sea-level states” or “anomalous situations” should be properly introduced. It is not a general term that is already defined elsewhere. In the manuscript, the term is used in two ways: (1.) as a term describing non-linear conditions with strong forcing, i.e., storm conditions and (2.) as situations with significant reconstruction error of the AANN method. But are these two definitions really synonymous? Reconstruction errors might also originate from the choice of the trainings data set (non-anomalous situations that were not considered during the training, stations with non-significant correlations), observational errors (drift in the data set), unconsidered variables (dependency on waves, etc.) and maybe more. Can these errors be excluded from the analysis?

Line 58-61, page 2: Comment: The paragraph starts with saying what you want to use AANN’s for and ends by saying that AANN’s are good in reproducing the data set they were trained for. Logically this is not fitting. I would connect this paragraph with the next one.

Another point is, that the AANN method is at least able to reconstruct weakly non-linear events. So, a failure of the method means that the ocean state should be highly non-linear. Is it possible to provide a classification for the non-linearity of the observed model state? At which point is an ocean state “non-linear” and at which point is it becoming “anomalous”?

Line 69, page 3: “NNs” instead of “NNss”

Line 80, page 3: What are “mean correlations”? The term has not been introduced properly.

2.Methods

I would suggest to split the chapter into two parts: Data and Methods, or to rename the chapter to “Data and Methods”. The chapter is dealing with the two aspects.

Figure1: page 4 (and all following figures): Please improve the quality of the figure in terms of resolution. Maybe you need also to restructure the figure to make it better readable. I can not read the names of the sea level stations and the titles in sub-figures (b-e). In the captions, the letters indicating the sub-figures should be presented together with the parameters that are shown. “Black bars” are rather “black lines”. How have the

water level residuals been calculated (that should be mentioned in the text)? I could not find this information in the manuscript.

Line 85-105, page 3-4: This paragraph belongs still to the introduction. It describes the model area and the general hydrodynamic of the North Sea, not the method that has been applied in the analysis. Instead, there should be an introduction into the use of the various data sets.

Line 109-114, page 4-5: It is unclear under which aspects the tide gauge observation stations were selected. Why were only southern and western North Sea tide gauge stations selected and not, for example, Danish stations?

More information about the observation data sets could be provided: data provider, monitoring period, frequency of the observations, data quality (amount of flagged data). Has the data been quality controlled? I know that at least some centers provide data to CMEMS without quality control. What about the time frequency of the data set? Is hourly data enough for the analysis, especially for the analysis of storm cases?

Line 115-133, page 5: It is unclear at this point why NEMO model data sets play a role in the analysis. There should be a short introduction into the scope of the study, explaining the different data sources and their role in the assessment.

The introduction of the NEMO model configuration could be extended as well. I'm for example not entirely sure if the model run on its own or if it run as a component of the GCOAST model system, and which other components of GCOAST were included as well. Information about tidal forcing (boundaries, tidal potential), the initial data set and the model spin-up and forecast period should be required as well. Did the model use nesting in the transition area between the North Sea and the Baltic Sea. 3.5 km seems to be a rather coarse spatial, horizontal resolution for this area. Has the quality of the model with regards to sea level predictions been analyzed? If yes, could the authors provide a reference to the model validation study.

Line 130, page 5: It is stated that ERA-5 atmospheric and wave parameter are provided by C3S. Have the wave parameters been used in this study? Otherwise, I would only mention the atmospheric product.

Line 134-145, page 5-6: The autoasssociative neural network method is introduced (chapter 2.2). However, the more general description (first paragraph, Line 135 to 145) requires some further explanation and re-formulation:

- The term “multivariate data” should be explained. I know the term only from data assimilation. I guess it is used here to point out that extensive, interdependent hydrographic data sets are measured that in the context of the AANN method are described as “multivariate data”.
- Another point is the definition of the sub-space of a given data space. What do you mean with “factors” (line 3, first paragraph of section 2.2)? Are you comparing the number of non-linear interaction processes between different monitoring parameters with the number of parameters? This is not so clear in the text. I think it needs further explanation or clarification. The way it is introduced here (at least the way I read it) is that the “sub-space” describes the reconstructed data set, not the original data set. So, errors in the reconstruction lead to situations when the original observed data sets are not an element of the reconstructed sub-space, in anomalous situations. Maybe this becomes clear after you have introduced the method, but its not so clear here.
- Near Line 140 (line 6, 1th paragraph of section 2.2) you speak about a “proxy”. What do you mean exactly by this?

Figure 2 (b): The way the fitting lines have been drawn in figure are a bit arbitrary. It is not so clear to me why not 6 neurons were chosen or 8 neurons. Is the method very sensitive to the number of bottleneck neurons?

In chapter 2.2 and sometimes in other parts of the document, the term “model” might be used in ambiguous ways. It usually refers to the “AANN model”, but might also refer to the “NEMO model”. I would suggest to always write the full name of the model.

In the manuscript, the method is introduced and the results are presented, but general information on how the neural network has been trained is not provided (see general points at the beginning of the review).

- **Anomalous sea-levels and their relationship with atmospheric conditions:**

This chapter falls into 2 parts, dealing with the analysis method (Line 200-235) and analysis of application cases (Line 240-280).

3.1 Analysis method:

The analysis method uses arbitrary limiters, which depend on the application at hand. The values for the threshold error, the minimum number of tide gauges and the minimum exceedance time of the threshold could be motivated with the AANN analysis results for the North Sea, either the entire testing year or the two application cases for 2 storms.

Line 200-205, page 8: I would use "sub-set of tide gauges" rather than "several" tide gauges.

Line 200-205, page 8: What is the given threshold error? It is not mentioned in the manuscript.

Line 220-225, page 9: What does "resemble" here mean? Are the graphs AANN_less and AANN_resid visually identical to AANN?

Line 220-225, page 9: I would not say that the model physics is constrained linear processes, which follow a Gaussian statistic. Especially turbulence is non-linear. One can say maybe that the model processes are not highly non-linear and that non-linearities are often dampened away by dissipation. It is also so, that interactions between processes can amplify nonlinearities in the model.

Line 220-225, page 9: I don't understand the second sentence. What does the observational data errors have to do with AANN_NEMO specifically? Are they not the same for all reconstruction models?

Line 220-225, page 9: You refer to Fig. 4c, which I can not read well.

Figure 4 (c): How have the water level residuals been calculated? Tidal Harmonic Analysis? How many components have been used? I can't see the figure well, due to its low resolution, but at some stations there seem to be still some harmonic signals. What about the reconstructed time series? Are the results of AANN_resid shown? Were the water level residuals of the NEMO model sea levels calculated? Maybe you are showing water level anomalies rather than residuals.

Line 230-235, page 9: It is not entirely clear which criterion was used, as the threshold error is undefined. What is the validation period that is analyzed in table 1?

Table 1: Which period does the analysis cover? How does the validation period relate to the training or testing period?

3.2 Application cases:

The analysis should be extended with a more detailed discussion of the results. Often the facts are presented without analyzing them in detail. The figures 5-8 seem to be rather separated from the discussion in the text. Sometimes a feature is singled out, like the sea level difference (observation-model) series at Lowestoft (fig. 5e), but all the other results in the figure 5e are not discussed and the results of the other sub-figures 5a-f are not discussed either. The same is true for figure 7 (Storm Burglind). The assessment is qualitative, not based on a statistical evaluation of model errors, especially water level peak errors. It seems also to be unclear, why the NEMO sea level prediction errors are discussed, as they do not seem to be directly related to the reconstruction error. For Storm Heinrich, for example, the reconstruction error is largest for stations in the southern North Sea (Dover, Oostende, Europlatform, Vlissingen), whereas the largest model error is occurring at Lowestoft. So, the conclusions are a bit uncertain.

I think the figures 5d and 6b,c,d and figure 7d and 8b,c,d for the two storm cases: Heinrich (June 2017) and Burglind (January 2018) are at the heart of the assessment. The analysis of the two cases shows different results in the reconstruction, which are briefly discussed, but could be presented in a more consistent way. It would be rather be interesting to analyze the ability of the hydrodynamic model to reproduce the observed anomalous correlation events figure 5d and 6d, as well as figure 7d and 8d.

One more comment: The exact dates for the two storm events should be given, as the names of storms vary from country to country.

Line 240-241, page 10: The reference to figure 5 seems to be incorrect. Are you sure that you don’t want to refer to figure 6b and 6c?

Line 240-255, page 10-11: Storm Heinrich: Removing the southern North Sea stations Dover, Oostende, Europlatform, Vlissingen leads to a better reconstruction of the sea level (Fig. 6c). Does this mean that the non-linear events with (anomalous spatial correlations) come from these 4 stations? The results: measured sea level minus reconstructed sea level (AANN) (Fig 5d) demonstrate the reconstruction error (related to anomalous spatial correlations) and Fig. 6c seems to proof this. Figure 6b seems to show that the reconstruction error is not related to the tides, which likely means that it is driven by meteorological forcing. NEMO seems to be able to reproduce some of the nonlinearity features at Oostende and Vilssingen, but not at Dover. Why is that the case? PCA seems to be able to catch the event, but it also shows an event for Den Helder. Why is that the case? Which method is having a problem here?

Line 255-280, page 11-13: Storm Burglind: Removing tides seems to lead to a negative reconstruction error at the 4 Southern North Sea stations: Dover, Oostende, Vlissingen (Fig. 8b). What does this mean? Is the trained AANN_resid method reproducing non-linearities that are not in the measured data set. Is this evaluated as a reconstruction error as well? In contrast, the AANN_resid method is generating a positive reconstruction error at Den Helder. Does this mean that atmospheric forcing is stronger in driving non-linearities there? Is this related to the training method? Removing the southern North Sea stations Dover, Oostende, Europlatform, Vlissingen leads to a reconstruction of the sea

level (Fig. 6c) with some positive error. Does this mean that the non-linearities were stronger in the central North Sea? Can this be proven by training an AANN that does not consider other stations? It seems that the separation into station (AANN_less) suits better the case of storm Heinrich. NEMO seems to be able to reproduce some of the nonlinearity features, which should be mentioned, but they are not as pronounced as the observed ones. Differences could be analyzed in detail.

Line 270-280, page 12-13: The reconstruction using PCA seems to work less well for Storm Burglind than it does for Storm Heinrich. Is this because only leading modes were considered? The number of PCA reconstruction modes is not mentioned in the manuscript. If this is the reason, then it should be made clearer in the text. Line 272: What is a false alarm?