

Evaluation of numerical models by FerryBox and Fixed Platform in-situ data in the southern North Sea. Haller, Janssen, Siddorn, Petersen & Dick.

General comments

The main thrust of this paper is a comparison of the temperature and salinity output from two numerical models with data from FerryBox tracks and an in-situ sampling station in the southern North Sea. The model outputs span different years but overlap in 2011-2012. A secondary thrust is demonstrating that the FerryBox data is of good quality. The paper presents a number of statistical measures to compare each model against the observations, both along ship tracks and at three locations, close to the English coast, in mid southern Bight, and close to the German coast where a fixed mooring platform provides in-situ observations.

Specific comments

The paper has a number of authors, and unfortunately it shows because different sections of the paper are clearly written in different styles with different levels of written English proficiency; it would benefit from a single author (i.e. Siddorn) acting as editor.

I came away from the paper with two messages 1) FerryBox observations are good measures of surface water mass properties, albeit with some issues associated with pipework locations, 2) both models generally agree offshore but both experience some difficulties in matching water properties near coasts, both English and German. The plethora of statistics is a bit overwhelming and could be trimmed back to the essence of the comparisons, the reporting of the statistics is rather monotonous with little insight into the reasons for the differences.

I struggled with the descriptions of the data extraction from the models, it isn't clear to me exactly what model data is used to compare with the FerryBox observations, what averaging is done (vertical and horizontal) to attempt to compare like-with-like.

The attempts to explain the differences between models and observations are confusing and sometimes contradictory. In different places in the manuscript vertical mixing in the models is given as a reason for disagreement, but sometimes this is described as *underestimating* mixing and sometimes as *overestimating* mixing. But no effort is made to explain properly what the mixing schemes are in the models or how their differences may lead to discrepancies. It is also a surprising result that AMM7, which includes data assimilation of surface temperatures does not show better results than BSH; it is pointed out in the text that the discrepancy between AMM7 and the MARNET observations is also surprising given better agreements in other (MyOcean) studies, but no effort is made to resolve these points.

I am puzzled by the choice of locations for comparisons ... point P2 is described as "marks the TD/TC meeting point in the Oyster Ground area" (caption figure 1) and this caption describes route TC (Copenhagen-Bergen, in green) but no route appears on the figure and no mention of TC occurs in the text.

I found the introduction section rather German-centric in its choice of references. Many well known UK, Belgian, Dutch and Norwegian references are ignored in preference to some rather obscure (or less well known) German references.

The model sections need improving, it wasn't made clear until deep in the results section that the BSH model is a nested grid with different fixed resolutions, there is an implication that the grid varies BETWEEN 900m and 5km. The surface and lateral boundary conditions need to be clearer, e.g. surface waves are mentioned – are these used at all? Which rivers have daily averaged data, and what kind of climatology is applied to the others (and how many?). AMM7 is said to assimilate SoO data ... where? Is this likely to impact on results in the southern North Sea? Satellite SST is also said to be assimilated, so how is this done? And what model data is actually used? These models are forecast models, are you using part of the forecast, initial conditions, or a specific hindcast?

Technical corrections (a selection)

Abstract: line 14 ... "Statistical errors differ between the models and the measured parameters, as the root mean square error (rmse) accounts for BSHcmod v4 to 0.92 K, for AMM7 only to 0.44 K. For salinity, BSHcmod is slightly better than AMM7 (0.98 and 1.1 psu, respectively)." Is poor English and needs rewriting

P357, line 10: "salinity near the coasts is only about 15–25" – near all coasts??

P357, line 13: "several factors like bathymetry, density distribution and wind stress" – I would add tides to this list.

P358, line 8: "installed on ships of opportunity (SoO), as well as on fixed onshore stations near harbours, river banks or estuaries." Examples of these other systems?

P358, line 13: "without limitation of energy", I think you mean "without power supply issues"?

P359, line 22: "Satellite imagery is somewhat limited regarding the time resolution and restricted to certain parameters. Moreover, satellite coverage is limited in coastal regions and in the vicinity of land" I don't believe this to be the case, so better justification for this statement needed.

P360, line 2: "The aim of this study ..." say "The aim of the present study..." to avoid confusion with the previous references Wehde and Petersen.

P360, line 6: "and identify the limitations and weaknesses of the operational models AMM7 and BSHcmod v4." Beyond the obvious that both models show less good agreement with the observations near coasts I think more effort is needed to explain these differences.

P360, line 25: "in case of severe errors" meaning what?

P361, line 15: Is all the technical detail really needed? A reference will do I think.

P362, line 3: "MARNET has a long tradition of monitoring" how long?

P363, line 12: "extrapolated from the lowest pressure level height" how high? Does it matter?

P363, line 28: "hybrid s-sigma terrain following coordinates are applied with 50 equally spaced levels" not sure I understand what is meant by this ... the purpose of using hybrid s-sigma is to allow some variability between levels, please explain.

P364, line 16: "the bias can then be described as the mean difference between simulations and observations" is incorrect given the following equation which shows the bias to be the difference between the mean of the observations and the mean of the simulation.

P365, line 20: "A value of 0.5 means that the model is on average 0.5 times the standard deviation off the observations. " What does this mean??"

P367, line 15-20: repeated text

P368 ... A number of things need to be made clearer about how the data are extracted and compared. One factor not explored is how well the models represent the tides because even 30-minute discrepancies in tidal phase could influence the temperature/salinity values extracted from the models.

P370, line 6: "The differences have been marked according to the double stde of the FerryBox data which has been described above" what? Where?

P370, line 12: spelling "spatial"

P371, line 6: "the differences as well as the according measures bias, stde and skvar show average values," what is meant here?

P371, line 27: "one could argue" one needs to demonstrate, not argue.

P375, line 17: "in some cases it may be right for the wrong reasons. Agreement is achieved when observed salinity happens to be in same range than tidally varying model values; otherwise there is no agreement." Rewrite this!

P376, line 9: "getting the annual cycle and the amplitude in the correct phase in their respective time period resulting in mainly good agreement" how can you tell this from Figure 10?

P377, line 6: "It could be argued, that a second drop has been there, but at a shifted position which could not be detected by the FerryBox." Pure speculation!

P377, line 10: "It is known for example that the AMM7 model underestimates flushing in the German Bight" reference?

P377, line 16: "freshwater eddies far away from the coasts" what are these? Where do they come from?

P378, line 26: "It is not clear why both models should both predict freshening in the summer 2011 which in fact did not occur." Perhaps a bit more investigation needed??"

P379, line 13: "The spectral densities of each time series are located in the same range". It would indeed be surprising if they didn't! what is not discussed is the SLOPE of these spectra, especially in the case of BSH.

P381, line 24: Is the Schulz-Stellenfleth & Stanev reference the correct one? (seems to deal with water levels not temperature and salinity).