Copenhagen, 23 August 2018

Dear editor,

We appreciate very much the comment from the two anonymous referees, which have been extremely useful for improving the manuscript. Below, you find the referees' comments with our responses in *italic*. Section numbers and line numbers refer to the revised manuscript.

Response to Anonymous Referee #1

1 General comments

The paper addresses a topical issue in the operational oceanography in the marginal seas – whether to introduce ensemble forecasting. The Authors have run the wave model WAM for the Baltic Sea with different horizontal and spectral resolutions and different atmospheric forcing to study whether one should increase resolution or introduce ensembles to provide better forecast accuracy. The question is interesting, but one would expect a more thorough and systematic approach in building and introducing the model and forecast system configurations and in analysing the results.

It has been long known, although not perhaps explicitly said, that the open sea areas of the Baltic Sea, where the shallow water effects can be neglected, do not much benefit of reducing the grid size. That said, there are several areas, where the high resolution is important to solve the shallow water effects and address the effects of islands and irregular shoreline on the wave fields.

> *It may be common knowledge that for open sea areas in the Baltic Sea, wave forecasts will not benefit from (further) decrease of the grid size, '..although perhaps not explicitly said'. It is not clear to us what the intention of this remark is; we find it a perfect motivation for our study. We have not been able to find any reference that states this, thus we have not been able to include this view. If the referee can help us in this matter, we would be happy to include such references.*
>
> .

Due to the small size of the Baltic Sea, the wave field is dominated by the wind waves and the accuracy of the wave forecast is largely dependent on the accuracy of the atmospheric forcing. Therefore comparing systems run with wind forcing from different NWP systems to address the question about choosing between ensembles and resolution is not entirely valid. Also the earlier studies the Authors refer to in the discussion most likely have different/older versions of the WAM model. Therefore the differences or non-differences cannot directly be connected to resolution or the atmospheric forcing. And also, if the time periods used in verification are relatively short (2-3 years) and different ones, the inter-annual variability in the wind conditions might also affect the accuracy. I'd expect more discussion about these subjects.

> *We agree that wind waves dominate the Baltic Sea and therefore comparing wave forecasts driven by wind forcing from different NWP-systems will not be entirely valid. However, due to an unfortunate error in (old) table 1, there has been a misunderstanding about the wind forcings in our study. We will explain this below and have revised the manuscript and in particular table 1 accordingly.*
>
> *Discussion of length of verification period has been included in discussion section. Btw., our three-years verification period is not short compared to other published work. For instance,* (Bunney and Saulter 2015) *use eight months,* (Pezzutto et al. 2016) *use seven months, and* (Cao et al. 2009) *use 12 months.*

The only driving factor in reducing the grid size in the open sea wave modelling is not the accuracy of the wave forecast. There might also be other factors. For example coupling of wave and 3D ocean models might benefit of having high enough resolution. Same applies also for atmosphere – wave coupling. Furthermore, the benefits of higher resolution come also when using high-resolution wind fields nowadays available for the Baltic (e.g. HARMONIE with 2.5 km resolution), which are not possible to get full benefit from if wave model resolution is coarser. I'd also like to see more discussion related to these subjects.

> *A discussion of higher resolution required by two-way coupling to atmospheric model and ocean model is certainly relevant. Similarily, we agree that increasing the horizontal resolution of the NWP-system may lead to better wind forecasts, in particular due to better descriptions of processes in extratropical cyclones, but we think this has already been mentioned in lines 52-58.*

Also, I think that the title should include indication, that you are focusing on the open sea, deep water areas.

> *We find our title catchy an fully covering our aim of the study, since we not a priori limited ourselves to the open sea areas. This was dictated during the work, due to available observations. Therefore, we have instead modified the conclusion and the abstracts and put more emphasis on the conclusion for the offshore sea.*

Is same wind forcing used both for HIGH and LOW NSB grids? This is not explicitly said in the manuscript. And is the forcing used the deterministic ECMWF or the HIRLAM wind field? Table 1 mentions both HIRLAM and ECMWF and Table 2 only states that one ensemble member is used as forcing, but not indicated whether the 1 ensemble is the ECMWF or HIRLAM deterministic forecast or something else. If HIRLAM is used for the HIGH and LOW NSB grids, then you are comparing wave model results with different NWP forcings against each other. Is it then question about resolution or different wind forcings? I suggest that you run both HIGH and LOW NSB grids using the control forecast from the ECMWF ENS system and compare the difference between them and the LOWENS to find what type of effects the resolution and introducing ensembles causes to the system. Furthermore, it is of course interesting to see, if with higher resolution wind forcing (e.g. HIRLAM) the results would further improve. If and when you use the HIRLAM forcing for the wave models, please specify, how you process the wind fields with 3 km resolution to the wave model grid with 5 km (and/or 10 km) resolution.

> *All configurations in the NSB-domain are forced with wind from the DMI-HIRLAM NWP-system. For the HIGH and LOW configurations the horizontal resolution is 3 km, while for the LOWENS it is 5 km. The ECMWF-forced North Atlantic domain is deterministic in all cases, and serves only as boundary data.*
>
> *Unfortunately, there was an error in in table 1 in the oroginal manuscript and this table has been corrected and extended in the revised manuscript.*
>
> *We have also included a sentence on how the 3/5 km HIRLAM wind fields are transformed to 5 and 10 km WAM grids by bilinear interpolation.*

Also, you should separately check, what can be addressed to spectral resolution and what to grid resolution. And also please check other parameters than SWH, for example it would be interesting to see, if there are affects to wave periods or directions, when using higher spectral resolution.

> *It could certainly be interesting to check the effect of changing spectral resolution and grid size separately, but would require a lot of additional work in the form of dedicated runs and analysis. The scope of the present work is to use different configurations already running in an operational environment to gain some knowledge regarding the resolution-ensemble issue. For the same reason, we can not isolate the effect from higher spectral resolution on wave periods and –directions.*

When looking through the supplement material, I was bit confused, why Arkona and Vahemadal were chosen to be the stations shown in the manuscript. E.g. looking Fig S2 Finngrundet, Nothern Baltic, Huvudskar show that HIGH gives lower rmse in many cases for the higher (of over 3 m) significant wave heights than LOW or LOWENS. If the lower rmse of LOWENS over longer forecast ranges come mainly from forecasting smaller than 3 m SWH it might not be that useful for duty forecasters. This type of conditions typically do not affect the marine traffic or the offshore structures, it is the extremes. Therefore it would be important to see how the different forecast systems behave in high wave conditions. The time period used in this study contains at least the January 2017 storm. It would be interesting to see a detailed comparison of the results during this storm and also in some other high wind events.

> *Vahemadal was chosen because it stands out against all other stations in the analysis (HIGH forecasts performs better for this station), and Arkona because for this station, together with Darss Sill, the ENSMEAN performed better than the other forecast classes, and this result is statistically significant (non-overlapping confidence bands). For the four last stations, the ENSMEAN perform best, but this result is not statistically significant. We have modified parts of (new) Section 6 the manuscript to make this clearer.*

Remember, however, that for SWH above 3 m there are few observations/forecasts and the statistics becomes very uncertain, as shown by the large error bars on Fig. S3 and S4, and also mentioned in the text.

Introducing the 11 January 2017 'Toini' storm as a case is a good suggestion. We have devoted a section to this in the revised manuscript, including and discussing the plot shown in Fig. 9 of SWH during January 2017, 48 hour forecast for Northern Baltic.

It is good that the Authors have shown that with ECMWF ENS forcing the accuracy of the wave forecasts is ok in the open sea areas of the Baltic Sea. I suggest that the authors do the more comprehensive model runs suggested above and also more detailed analysis of the results and also discuss the advantages and disadvantages of each system more thoroughly. Furthermore, it would be interesting to see how much skill the ensemble forecasts have for longer forecast period. To my experience, there is not much spread in the ensembles for the first two or three days and the true benefits of the ensemble system and probabilistic forecast usually comes with longer forecast ranges. It would be interesting to see up to which forecast lengths the ensemble system shows skill in forecasting the Baltic Sea wave conditions both in average and extreme conditions.

As explained above, our runs are all forced with DMI-HIRLAM ensembles, and therefore they only reach 48 hour forecast time.

The experience that the benefits of an ensemble system shows up after three days only may be true for ECMWF-ENS, but for the DMI-HIRLAM ensemble we see an effect for some stations already from about 36 hours (see Fig. S2)

Please also see my specific comments given below.

2 Specific comments

2.1 Introduction

Lines 29-35: I'd expect that the concept of deep and shallow water waves is introduced here, since this is one of the key issues in the discussion of the results.

Agree, we have done this.

Line 33: Bathymetry is important only if waves interact with bottom.

Agree, we have introduced a more precise formulation

Lines 29-25: How about weak non-linear wave-wave interactions?

> *Agree, (Non-linear) wave-wave interaction has been mentioned here*

Lines 41-42: Seasonal ice conditions vary quite a lot in the Baltic. Perhaps this description refers to an average ice winter?

> *Yes, this is included to remind the reader that sea ice is an issue in the Baltic Sea.*

Line 51-52: Is Baltic Sea shallow considering the average wave conditions? If then the use of higher resolution should make a difference, which is not in agreement with the conclusion drawn by the Authors later on. Baltic Sea is shallow compared to the Oceans, but when considering the typical wave periods/lengths in the Baltic, in most cases waves in the open sea areas are deep water waves, expect for high and extreme wave conditions.

> *The Baltic Sea may be 'shallow' compared to the world ocean, but is in general not shallow as regard wave conditions. We have therefore deleted the misleading sentence.*

2.2 Model and setup

This section needs restructuring. All information needed is basically given, but the order of things and the fact that some information is only given in Tables and the table is not referred in the corresponding place in text makes it difficult to follow.

> *We have re-structured section 2. incl. the tables to make it more understandable.*

Also please define explicitly, which wind forcing is used for LOW and HIGH configurations. Table 3 indicates that deterministic ECMWF forcing is used for the coarse, larger domain and HIRLAM (and possibly also ECMWF?) for the smaller high resolution domain.It is not clear to me if this Table refers to DMI operational setup or for the setup used in this paper.

> *We have corrected errors in Tables 1 and 2 regarding the wind forcing in the different configurations, as explained above.*

Lines 73-77: Please specify the source terms and formulations used in the model runs.

> *Source terms are described on lines 105-110, so we do not understand the referee's remark.*

Lines 78-82: Specify the horizontal resolution of the areas already here or cite a Table where they are given. I also suggest adding the resolution info to Table 1.

Horizontal resolutions are now given in Table 2.

Line 88: Specify the various sources used to compile the bathymetry

These are now listed in Table 2.

Line 121-122: You use only 11 members of the total 50 available from ECMWF. How do you select, which members you use?

As explained above, we use DMI-HIRLAM atmospheric ensemble forcing, of which a subset of 11 ensemble members are run at routine basis at DMI. The ensembles are generated by perturbing a number of processes, e.g. cloud physics, which do not have a direct impact on the wind field. A subset of 11 ensembles was recommended by the DMI-HIRLAM ensemble developers to cover the spread in surface wind.

Table 2: It is unclear to me what the column 'Ensemble members' mean for LOW ad HIGH.

'1' meant deterministic forecast. We have replaced by '-' in Table 2

2.3 Observations

Why not use Helsinki wave buoy data from Gulf of Finland? This should be available through CMEMS. Helsinki site mostly represent deep water conditions and it would be interesting to see, how the setups behave there compared to Vahemadal.

Helsinki wave buoy does not have many valid data in our verification period, see plot below. Therefore it is not included.

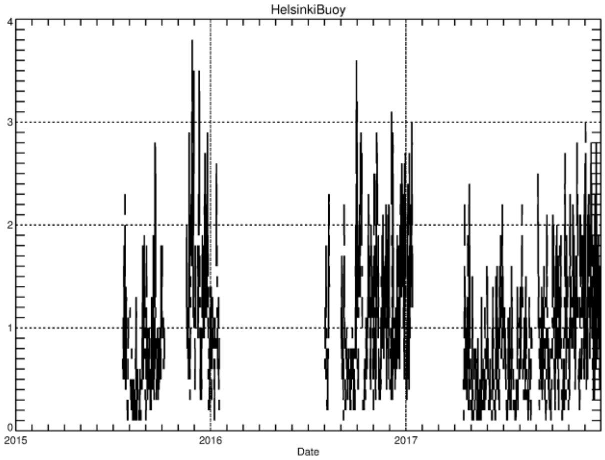


Table 3 gives only model depth at the buoy locations. It would be important to know also the actual depth at the buoy locations to evaluate, whether the model is adequately able to account for the deep and shallow water features in the wave field.

The actual water depths have been listed in Table 3.

It is bit unclear to me, what is the function of Figure 3. The details are lost here, since the images are so small. If they area meant to represent the overall description of the wave conditions at each site, please also (or maybe instead) give some description in the text. And if it is to show the gaps in the measured data, that could be put in a table.

We have moved Figure 3 to the supplementary information; we think the reader should have the opportunity to see which data series were actually used. To give the reader a feeling of the data, we have in the main text replaced figure 3 by a new table 4 showing frequency of swh-height in different intervals.

2.4 Verification

Give some explanation, why you have selected Bothnian Sea, Arkona and Vahemadal stations for more detailed analysis

See above

I also suggest doing some verification of the forcing wind fields.

We have included wind verification for two Danish coastal stations in a new Section 5.

In addition to verifying the general accuracy, I'd expect to see some verification of high wind/wave events. They are the most important ones to forecast accurately considering the marine traffic and offshore structures.

> *Sections 5.1.1 (RMSE and BIAS as function of SWH) and 5.3 (Brier score verification) are already verifications of high wave events. In addition we have included an intercomparison of forecasts of the Jan 2017 Toini storm.*

I would also expect more discussion of the importance of wind field accuracy on the accuracy of the wave forecast. The accuracy of wave forecast in the open sea areas might not benefit from higher resolution in the wave model grid, but what about when the wind forcing has high resolution, such as the HARMONIE forecasts run for the Baltic with 2.5 km resolution in several of the MET services. In order to account for the benefits of this, higher resolution in wave model grid might become important.

> *This point is addressed at the end in the Introduction.*

2.5 Discussion

You should very carefully analyse and explain, what you are actually comparing in Table 2. To my understanding you are comparing wave forecast system, which have different resolutions, wind forcing and also most likely different wave model versions. So the differences in accuracy cannot solely be attributed to resolution.

> *We think the referee refers to table 4 (present table 6). In that case we agree and have put in an appropriate sentence in the discusion*

Table 4 – Why have you not calculated rms errors for the Helsinki wave buoy for LOW, HIGH and HIGHENS?

> *Due to lack of observations, see above.*

I'm not sure why the ice coverage is discussed here. You are comparing the forecasts against buoy measurements and the buoys are recovered well before there is a risk of ice in the area. Therefore handling of ice should not cause any problems in your verification results. That said, you of course have this element during the season and in the areas where you are unable to do the verification. You could also give a short description of the ice conditions in 2015-2017 so that readers would be able to evaluate, how big effect this might be.

> *The effect of ice cover is described because it is a potential systematic error source of the wave forecasts, and in addition may introduce a bias in the verification due to withdrawal of observing*

buys during winter time with high waves. We have extended the discussing, ending up with making it likely that ‘sea ice problems’ are small during the period 2015-2017.

Response to Anonymous Referee #2

OVERVIEW:
This paper will serve a useful purpose in documenting performance of an operational wave forecast modelling system for the Baltic and in assessing and discussing the relative benefits of increasing wave model resolution versus a probabilistic forecast system in this specific scenario - where approximately an order magnitude increase in computing power has been available. For such a study, the authors have done a good job with being concise in their use of probabilistic verification metrics and delivered a clear set of results.

However, I would recommend that publication is made subject to a number of major revisions. These are required in order to address a number of questions raised by the study, but which the authors have only dealt with very briefly or passed over:

1. For a wind driven wave model, the nature and quality of the forcing winds are a key consideration in the model performance. The driving wind model therefore needs to be well documented and, specifically for this paper, any differences in horizontal resolution associated with the deterministic and ensemble forecasts systems need to be provided clearly in Section 2 These were not clear to me on my read through, and I am left with the impression that the authors have compared a 10km wave model with a 5km wave model but using a similarly specified wind model for both deterministic and ensemble forecasts?

> *The HIGH and LOW deterministic models are forced with wind from 3 km DMI-HIRLAM, while LOWENS are forced with wind from 5 km ensemble DMI-HIRLAM. In our opinion, this disfavours the LOWENS forecasts, since the ensemble runs are forced by lower-resolution wind fields. We have modified Section 2 of the manuscript to make clearer which wind forcing is used for the different configurations.*

2. If this is indeed the case, then I think the wind forcing being used, wave model resolutions chosen and available observations naturally lean the study toward favouring the ensemble. This is acceptable, but needs to be acknowledged and discussed further within the paper. From a wind perspective, if no higher resolution atmosphere model that will improve representation of the land-sea boundary layer is available then the ensemble's provision of multiple answers will generally help the verification scores from that system. Whilst a costly enhancement, the change from (LOW) 10km to (HIGH) 5km resolution may not be enough to significantly enhance wave forecast performance in the coastal zone and, besides, only one observation site is available to illustrate coastal performance. This means that it is difficult for the reader to get a clear picture of what advantages the HIGH res model is expected to yield - I'd suggest that might be improved by some visualization of model fields in order that the impact of changes from LOW to HIGH over the wider region can at least be illustrated.

> *We do not understand why the referee thinks that our configurations favours the ensemble. In our opinion, the setup disfavours the ensemble, as pointed to under 1.*
>
> *The Land-sea boundary problem has been accounted for using water-wind, as described in the manuscript.*

To illustrate the performance of the different configurations, we have included in Section 2 forecast fields valid at the peak of 'Toini' storm, January 2017.

3. Although, in my view, the experiments favour the ensemble system, the paper still raises a valid point: which is that when using regular grid wave models and an order of magnitude computing resource to invest then the ensemble will likely provide a better return, in terms of improving forecast skill over the larger offshore part of the domain. However, in order to make this point the authors also need to be mindful of and discuss the study within the context of rather more of the open literature than they have done. For example, Cavaleri et al (2018) provide an exhaustive discussion of coastal processes and how wind and wave models need to improve in order to properly represent these - it would be good if the authors can set out where and how the HIGH system attempts to address these aspects of coastal forecasting better than the LOW or LOWENS systems. Similarly, there is also the question of whether an unstructured or refined grid approach would enable significant improvements in coastal regions of the domain whilst keeping the model efficient offshore and enabling a best of both worlds approach (e.g. Bunney and Saulter„ 2016). So I would recommend that the authors try to address these aspects of the paper with appropriate references in both Sections 1 and 6.

Thanks for pointing to these aspects. We have included a short summary of these aspects in the conclusion section.

SPECIFIC COMMENTS

Paragraphs at line 44 and 48. I think this discussion could be a bit more expansive? The authors have followed through the practical viewpoint where the wave model is scaled to the NWP and then resolution is increased if there is spare resource. This is a quite standard 'in practise' way of working, but as a motivating point it would be good if the authors could expand on what scales they believe are required for an idealised/ pragmatic wave forecasting system that dealt with both coastal and offshore areas of the region.

We do not believe that there is one simple model setup, performing well both for offshore and coastal conditions. One has to resort to nested configurations, unstructured mesh, or subgrid-scale representations. Instead of in the introduction, we have incorporated these aspects in the conclusion section (see above).

Sentence at line 64. I'm not convinced that the ensemble vs resolution increase argument is generic, rather it depends on where the model is being used and how end-users will deal with the resulting products. So I think it would be better to contextualise this argument to the situation in question - a wind-wave dominated regional sea with a mixture of offshore and coastal regimes.

We do not quite understand this point. At line 64 we mention two ways to spend additional computer resources: ensembles or increased resolution. We do not bring forward any arguments or analysis at this point.

Sentence at line 112. I'm not convinced the information about the spin-up is that useful.

OK, we have removed it.

Paragraph at line 117. Around here would be an excellent place to add further detail regarding the NWP forcing.

We have added info on wind resolution to table 2.

Paragraphs at lines 247 and 255. The dependencies of RMSE/bias on SWH are to be expected when matching up deterministic forecasts since small timing errors in the predicted wave time-series will have larger impacts on the model-observation match-up in the upper percentiles of the SWH pdf than in the lower percentiles.

Yes, this is now mentioned in the conclusion

Section 6.1. It would be useful to state the resolution of the NWP systems underpinning Tuomi et al.'s wave models.

We have added that info to (present) Table 6.

Section 6. For completeness it would be worth discussing the spread-skill characteristics of the LOWENS system. At the sort of short forecast ranges discussed, these systems are usually under-spread and it would be useful to know if this is also the case here (and if not, why not?). The ability of the ensemble to properly generate spread provides the difference between running a system that provides some improvements to forecast verification vs a deterministic model through a partial sampling of forecast uncertainty, and one that genuinely samples the likely observed outcomes.

We have included rank histograms as Section 6.4 and discuss these.

Section 6. This would be a good place to talk through the computational limits placed by using a regular grid scheme in this region and some of the other modelling options that might allow some best of both worlds solution to be achieved in future. Its fair to say that in supercomputing terms a resource increase of order 3-10 times might be the maximum expected over 1 or 2 new systems, so the problem highlighted here is important.

We think we have already replied to the issues of nested domains, unstructured grids etc. above.

Better Baltic Sea wave forecasts: Improving resolution or introducing ensembles?

Torben Schmith, Jacob Woge Nielsen, Till Andreas Soya Rasmussen, Henrik Feddersen

Danish Meteorological Institute, Copenhagen, Denmark

Correspondence to: Torben Schmith (ts@dmi.dk)

Abstract. The performance of short-range operational forecasts of significant wave height in the Baltic Sea ~~in three different configurations~~ is evaluated. Forecasts produced by a base configuration are inter-compared with forecasts from two improved configurations: one with improved horizontal and spectral resolution and one with ensembles representing uncertainties in the physics of the forcing wind field and the initial conditions of this field. Both the improved forecast classes represent an almost equal increase in computational costs. The inter-comparison therefore addresses the question: would more computer resources most favorably be spent on enhancing the spatial and spectral resolution or, alternatively, on introducing ensembles? The inter-comparison is based on comparisons with hourly observations of significant wave height from seven observation sites in the Baltic Sea during the three-year period 2015-2017. We conclude that for most ~~stations~~wave measurement sites, the introduction of ensembles enhances the overall performance of the forecasts, whereas increasing the horizontal and spectral resolution does not. These ~~stations~~ sites represent offshore conditions, well exposed from all directions with a large distance to the nearest coast and with a large water depth. Therefore, the detailed shoreline and bathymetry is also a priori not expected to have any impact. Only for one ~~station~~site, we find that increasing the horizontal and spectral resolution significantly improved the forecasts. This ~~station~~ site is situated in nearshore conditions, close to land, with a nearby island and therefore shielded from many directions. This study therefore concludes that to improve wave forecasts in offshore areas, ensembles should be introduced~~,~~. For near shore areas, the study suggests that additional computational resources should be used to increase the resolution.~~while for nearshore areas better resolution may improve results~~.

1 Introduction

Severe surface waves affect ship navigation, offshore activities and risk management in coastal areas. Therefore, reliable forecasts of wave conditions are important for ship routing and planning purposes when constructing, maintaining and operating offshore facilities, such as wind farms and oil installations.

Waves are generated by energy transfer from surface winds that act on the sea. The energy transfer is determined~~The development of waves is further influenced~~ by the *fetch* (the distance, over which the wind acts), and by the *duration* of the wind. For *deep water waves*, defined as the wave height being much smaller than the water depth, d~~D~~issipation of the wave energy occurs through internal dissipation mainly~~,~~. For *shallow water waves,* defined as the wave height being comparable to the water depth,~~-~~ dissipation through bottom friction and through wave breaking over a shallow and sloping sea bed becomes important. Shallow water waves may also be refracted over a varying bathymetry ~~Thus~~Therefore, a correct and detailed description of the bathymetry is important for correctly forecasting waves in coastal areas and other shallow sea areas.~~ Also refraction of waves is influenced by the bathymetry.~~ Other factors~~, which~~ with a potential~~ly has an~~ effect on the development of waves include~~s~~ nonlinear wave-wave interaction, ocean currents, time-varying water depth due to variations in sea level, and sea ice coverage.

The Baltic Sea is connected to the world ocean through the ~~Transition Area~~Danish waters with~~ the~~ shallow and narrow ~~Danish ~~Straits (see Figure 1), and this allows virtually no external wave energy to be propagated into the area. The Baltic Sea consists of a number of basins with depths exceeding 100 m, separated by sills and~~ shallow~~ water areas with more moderate water depths. Between Finland and Sweden lies an archipelago with complicated bathymetry on very small spatial scales. The wind is in general westerly over the area, and the most prominent cause for severe wind and wave conditions is ~~lows ~~low pressure systems passing eastward over central Scandinavia. Winter ice occurs in the northern and eastern parts of the Baltic Sea. There is no noticeable tidal amplitude or permanent current systems.

Short-term forecasting of surface waves is done by a wave model, forced with forecasted wind from an atmospheric numerical weather prediction (NWP) model. The equations of the NWP model are discretized on a horizontal grid with a certain spatial resolution, which determines the maximum spatial resolution of the wave model. ~~Due to limited~~The available computer resources~~, only~~ put a limit on ~~certain ~~the horizontal grid spacing, which can be afforded.

Technical development has increased the computational resources, making possible to increase ~~With additional computer resources becoming available, ~~the horizontal spatial resolution of the NWP and wave models~~can be increased~~. This allows for an improved description and forecasting of the synoptic and mesoscale atmospheric systems, including the details of the associated wind field~~, by the NWP model~~. In addition, a more detailed description of the bathymetry improves the correct description of dissipation and refraction of waves, as argued above.~~ This is in particular true in shallow seas, such as the Baltic See.~~ Additional computer resources may also be used to improve the spectral resolution in the wave model. This includes the directional resolution and the number of frequencies included.

~~Historically,~~Increasing computer resources have ~~increased through time, and this development is expected to continue in future. This has~~also made~~ a development towards~~ ensemble ~~weather forecasts~~NWP possible. The purpose of ensemble forecasts is to improve forecast skill by taking both the initial error of the forecast and the uncertainty of the model physics into account. Furthermore, ensemble forecast allows for probabilistic forecasts, identified as a priority for operational oceanography (She et al., 2016), and allows for quantifying forecast uncertainty. Ensemble wave forecast systems have been implemented at global scale (Alves et al., 2013; Cao et al., 2009; Saetra and Bidlot, 2002) and more regionally in the Norwegian Sea (Carrasco and Saetra, 2008), and in the German Bight and Western Baltic (Behrens, 2015).

From the above discussion it is evident that additional computer resources can be used in different ways to change the wave forecast setup, in order to increase the forecast quality. The purpose of the present study is to investigate the effect on forecast quality of increasing the horizontal resolution and the spectral resolution vs. introducing ensemble forecasts. This will be done by verifying the DMI operational forecasting of wave conditions in the Baltic Sea in different configurations against available observations of significant wave height.

It should be mentioned that improving wave forecasts is not the only driving factor in reducing the grid size of the wave model. Coupling the wave model with atmosphere or ocean circulation models may give a better description of vertical fluxes of heat and momentum (Cavaleri et al., 2012). For instance, Alari et al.(2016) documented a significant improvement of modelled SSTs by the NEMO circulation model in the Baltic Sea when a two-way coupling to the wave model WAM was introduced. Doing such couplings may demand a high horizontal resolution to describe the fluxes most satisfactorily.

Also increasing the horizontal resolution of the NWP-system may lead to improved wind forecasts, due to in particular better descriptions of processes in extratropical cyclones. In these cases, where the wind field is strong and varying on a small spatial scale, also wave forecasts may be improved by running the wave model in a similarly high resolution.

This paper is arranged as follows. Section 2 describes the model and setup, ~~and~~ Section 3 describes the observations used and~~. T~~ the verification methodology is described in Section 4. Verification of DMI-HIRLAM wind forecasts is in Section 5, and the SWH forecast verification is in ~~and applied in~~ Section 6~~5~~. Results of the verification are discussed in Section 7~~6~~ and conclusions made in Section 8~~7~~.

2 Model and setup

The DMI operational wave forecasting system DMI-WAM uses the 3rd generation spectral wave model WAM Cycle4.5 (Günther et al., 1992) forced by the regional NWP model DMI-HIRLAM and the global NWP model ECMWF~~/~~-GLM. WAM Cycle4.5 solves the spectral wave equation, and calculates the wave energy as a function of position, time, wave period and direction. Derived variables, such as the significant wave height (SWH), are calculated as suitable integrals of the wave energy spectrum.

The DMI-WAM~~ suite~~ forecasts waves in a larger area than the Baltic Sea and therefore has a setup with two nested spatial domains of different geographical extent~~ and spatial resolution~~ (see Figure 1): North Atlantic (NA) and North Sea/Baltic Sea (NSB), of which forecast results from the NSB-domain are ~~used~~ analyzed in this study. The ~~North Atlantic~~NA domain uses the JONSWAP wave spectrum for fully developed wind-sea (Hasselmann et al., 1973) along open model boundaries, while the NSB domain use modeled wave spectra from the NA domain at its open boundaries (one-way nesting).

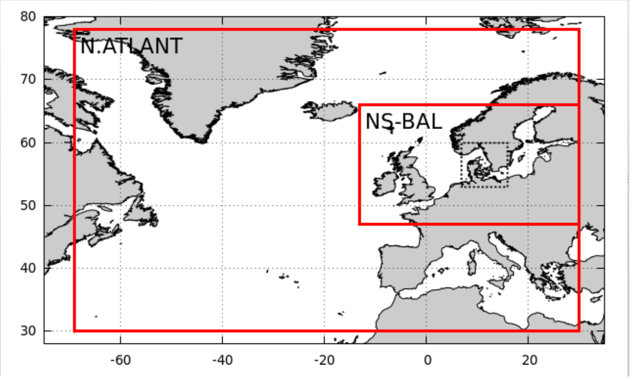


Figure 1 Nesting of domains in DMI-WAM. Outer frame is North Atlantic (NA) domain, inner frame is the North Sea/Baltic Sea(NSB)-domain. Dotted frame is the Transition Area. Only data from the NSB-domain are analyzed in this study.

~~In the North Atlantic domain, bathymetry is taken from Rtopo 30"×30" global bathymetry , while in the NSB domain, Rtopo is combined with local depth information from various sources, in order to obtain a more accurate bathymetry.~~

The wave energy is discretized into a number of wave directions and frequencies. To facilitate wave growth from calm sea, a lower limit is applied to the spectral energy. The resulting surface roughness parameterizes the effect of capillary waves, and corresponds to a minimum significant wave height of 7 cm.

The energy source is the surface wind. The sink terms are wave energy dissipation through wave breaking (white capping), wave breaking in shallow areas, and friction against the sea bed. Depth-induced wave breaking (Battjes and Janssen, 1978) is used in the NSB domain only, since in the ~~North Atlantic~~NA domain, the depth maps are not detailed enough for activation of this effect. The wave energy is redistributed spatially by wave propagation and depth refraction, and spectrally by non-linear wave-wave interaction. Interaction with ocean currents and effects due to varying sea level caused by tides or storms are not incorporated.

In addition to a land mask, we have a time-varying ice mask. Below ice 30% concentration, sea ice is assumed to have no effect. Above 30% ice concentration, no wave energy is generated or propagated, i.e. the effect is like that of land. The applied sea ice concentrations originate from OSISAF (http://osisaf.met.no/p/ice/) with a frequency of 24 hours and around 25 km true horizontal resolution, gridded to ~10 km horizontal resolution and interpolated to the WAM-grid. The ice cover is initialized every day at 00z, and kept constant throughout each forecast run. ~~The wave energy is completely dissipated in areas with sea ice cover above 30%.~~~~The ice concentration originates from OSISAF (http://osisaf.met.no/p/ice/) with a frequency of 24 hours and around 25 km true horizontal resolution, gridded to ~10 km horizontal resolution and interpolated to the WAM-grid. The ice cover is kept constant through each forecast run.~~

The surface wind forcing is provided by different atmospheric models for the two domains. For the ~~North Atlantic~~NA domain, wind is provided by the ~~ECMWF~~ ECMWF-HRES global weather forecast ~~in 16 km resolution~~ every 3 hours. For the NSB domain, the surface wind is provided every hour by DMI-HIRLAM~~, version SKA (3 km resolution)~~. Setup details are summarized in Table 1 ~~every hour.~~

~~To diminish coastal effects, DMI-WAM uses a special *water-wind*, in which the surface roughness everywhere is assumed to be that of water. This enhances the wind speed in the coastal zone, most important in semi-enclosed areas (bays, fjords, etc.). It is basically a way to sharpen the land/sea boundary, reducing influence of land roughness on near-shore winds. Setup details are summarized in Table 1.~~

Table 1 Specifications of DMI-WAM nested setup.

Domain	North Atlantic	North Sea/Baltic Sea
Longitude	69W-30E	13W-30E
Latitude	30N-78N	47N-66N
Atmospheric forcing	~~GLM~~ECMWF-HRES	*DMI-HIRLAM*
Boundary condition	JONSWAP	One-way n~~N~~ested
~~Bathymetry~~	~~Rtopo~~	~~Rtopo/IOW/GEO~~
Depth-induced wave breaking	No	Yes

~~The ice concentration originates from OSISAF (http://osisaf.met.no/p/ice/) with a frequency of 24 hours and around 25 km true horizontal resolution, gridded to ~10 km horizontal resolution and interpolated to the WAM-grid. The ice cover is kept constant through each forecast run.~~

~~DMI-WAM is cold-started once and for all using fully developed sea with a constant fetch of 30 km based on the JONSWAP spectrum. Subsequent~~ Each ~~model~~ forecast runs ~~are~~is initialized using the sea state at analysis time, calculated by the previous run as a six hour forecast~~. The first two weeks after cold-start is regarded as spin-up~~. The operational DMI-WAM suite is run four times a day to 48 h forecast range. Spatial fields of forecasted SWH and other variables are output in hourly time resolution.

Historically, ~~Three~~ three different configurations of the DMI-WAM setup have been ~~applied~~used, and data from these for the period 2015-2017 is the basis for the present verification. In the old LOW configuration, ~~the NSB-domain has approximately 10 km~~ the horizontal resolution is around 50 km in the NA domain and around 10 km in the NSB domain, and the wave energy is resolved in 24 directions and at 32 frequencies, corresponding to wave periods of 1.25-23.94 s and wave lengths of 2.4-895 m (in deep water). Bathymetry is ETOPO (Amante and Eakins, 2009) in the NA domain, and the Baltic bathymetry from IOW (https://www.io-warnemuende.de/topography-of-the-baltic-sea.html) supplemented by depth data from the Danish Geodata Agency (DGA) in the NSB domain. ~~An~~ More recently, an ensemble configuration (LOWENS) has been introduced ~~has~~ with characteristics identical to LOW, but with parallel run of 11 ensemble members forced with perturbed atmospheric fields (initial conditions and physics). Finally, in the also recently introduced HIGH configuration, the horizontal resolution is around 25 km in the NA domain and ~~approximately~~ around 5 km in the NSB domain, and the wave energy resolved in 36 directions and 35 frequencies, corresponding to wave periods of 0.94-23.94 s, and wave lengths of 1.37-895 m (in deep water). Bathymetry is RTopo (Schaffer et al., 2016).

All configurations are forced by winds from ECMWF-HRES in the NA domain and DMI-HIRLAM in the NSB domain. In the NSB domain, the LOW and HIGH are forced by the S03 version (3 km horizontal resolution), while LOWENS is forced by the S05 version (5 km horizontal resolution). The DMI-HIRLAM winds are interpolated to the WAM grids by bilinear interpolation. To diminish coastal effects, DMI-HIRLAM delivers a special *water-wind* to DMI-WAM, in which the surface roughness everywhere is assumed to be that of water. This enhances the wind speed in the coastal zone, most important in semi-enclosed areas (bays, fjords, etc.). It is basically a way to sharpen the land/sea boundary, reducing influence of land roughness on near-shore winds.

An overview of the DMI-WAM configurations is provided in Table 2~~Table 2~~.

Table 2~~2~~ Details of DMI-WAM configuration used in this study.

	DMI-WAM Horizontal resolution [km]		# wave directions	# wave spectral frequencies	Bathymetry		Atmospheric horizontal resolution [km]		Ensemble members	
	North Atlantic	NSB			North Atlantic	NSB	North Atlantic (ECMWF)	NSB (DMI-HIRLAM)	North Atlantic	NSB
LOW	50	10	24	32	ETOPO	IOW/DGA	16	3	~~1~~-	-
LOWENS	50	10	24	32	ETOPO	IOW/DGA	16	5	~~11~~-	11
HIGH	25	5	36	35	RTopo	RTopo	16	3	~~1~~-	-

When replacing the LOW forecast ~~setup~~ configuration with the HIGH configuration~~setup~~, the required computational resources for running DMI-WAM are increased by a factor of 2^2 (increase in horizontal resolution) × 1.75 (effective decrease in time step) × 1.5 (increase of number of directions) × 35/32 (increase of number of spectral frequencies) ≈ 11.5~~,~~. From the LOW to the LOWENS configuration, ~~while~~ it is increased by a factor of 11 (number of ensemble members~~s~~) ~~from the LOW to the LOWENS setup~~. Since these increases in computational effort are very similar, an intercomparison can contribute to answering the question: should additional computer resources be used for increasing the spatial and spectral resolution, or for sampling the uncertainty in meteorological conditions using ensembles.

The LOW and HIGH configurations both produce a class of deterministic forecast, which are also named LOW and HIGH, respectively. The LOWENS configuration produces a class of probabilistic forecast, called LOWENS. In addition, the ensemble mean defines a class of deterministic forecasts, called LOWENSMEAN.

To illustrate differences to be expected among the deterministic forecasts, we show 48 h forecasts of SWH valid at the peak of the 'Toini' storm on 10 January 2017.

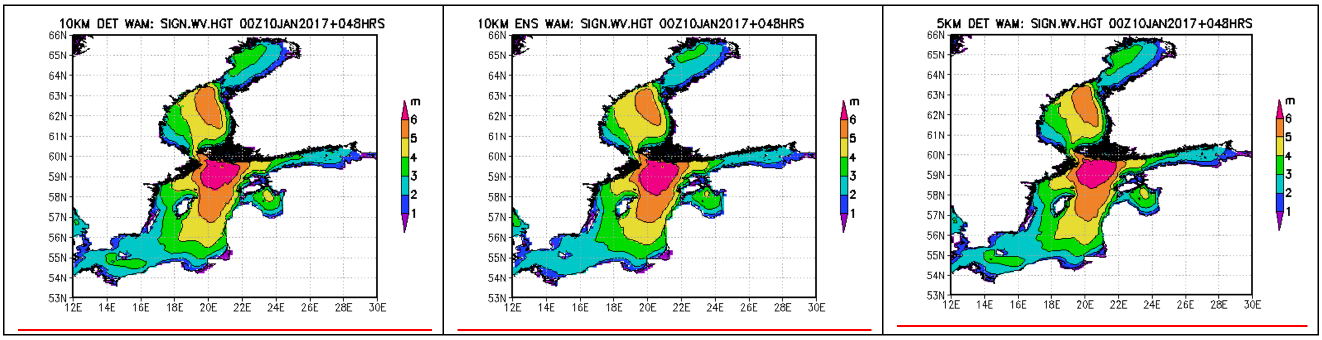

10KM DET WAM: SIGN.WV.HGT 00Z10JAN2017+048HRS
10KM ENS WAM: SIGN.WV.HGT 00Z10JAN2017+048HRS
5KM DET WAM: SIGN.WV.HGT 00Z10JAN2017+048HRS
66N
65N
64N
63N
62N
61N
60N
59N
58N
57N
56N
55N
54N
53N
12E
14E
16E
18E
20E
22E
24E
26E
28E
30E
m
6
5
4
3
2
1

Figure 2 Forecasted (48h) SWH at the peak of 'Toini' storm 10 january 2017 00z for LOW (left), LOWENSMEAN (middle) and HIGH (right) forecasts.

All three forecasts agree in the gross features of the forecasted SWH field. However, there are differences, e.g., northeast of the island of Gotland, the area with SWH above 6 m extend further southward in the LOWNSMEAN forecast, than in the LOW and HIGH forecasts.

3 Observations

Observed series of SWH from ~~stations~~ wave measurement sites in the Baltic Sea obtained from the Copernicus Marine Environmental Monitoring System (CMEMS) database are used. None of the series has a continuous record over the three-year period 2015 – 2017. Data gaps may be due to malfunction, maintenance or withdrawal of the instrument. The latter occur during winter due to the possibility of ice. We selected ~~station~~sites with valid observations that covered more than 40% and were distributed reasonably throughout the study period. Figure 3~~Figure 2~~ and Table 3~~Table 3~~ show the positions and water depths of the wave ~~selected~~ measurement ~~station~~sites together with the bathymetry of the Baltic Sea. Some ~~station~~sites did not observe at the full hour. Observations from these ~~station~~sites were ascribed to the nearest full hour, if the time distance between the observation time and the full hour was less than 15 min, otherwise not used. All observation series used are shown in ~~Figure 3~~Figure S1. The frequency of observed SWH in different intervals for each site is given in Table 4

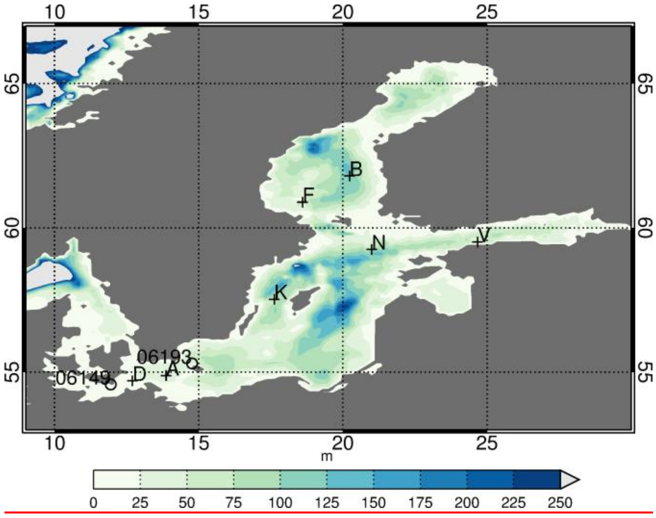


Figure 3 Map of the Baltic Sea with bathymetry and positions of wave measurement ~~station~~ sites marked with crosses. For details about ~~station~~sites, see Table 3~~Table 3~~. Meterological stations used in the wind verification of DMI-HIRLAM are marked with circles.

Table 3 Details of ~~stations~~wave measurement sites.

Observation site	Lon	Lat	~~Model d~~Depth [m]	
			Model	Actual
A Arkona WR	13.9	54.9	46	45
B Bothnian Sea	20.2	61.8	118	~120
D Darsser Sill WR	12.7	54.7	20	21
F Finngrundet WR	18.6	60.9	56	67
K Knolls Grund	17.6	57.5	63	90
N Northern Baltic	21.0	59.2	68	~100
V Vahemadal	24.7	59.5	18	5

Table 4 Observed frequency of SWH in different bins for wave measurement sites.

SWH [m]	0-1	1-2	2-3	3-4	4-5
Arkona WR	0.47	0.39	0.12	0.01	0.00
Bothnian Sea	0.46	0.38	0.12	0.02	0.01
Darsser Sill WR	0.67	0.31	0.02	0.00	0.00
Finngrundet WR	0.69	0.27	0.04	0.01	0.00
Knolls Grund	0.62	0.31	0.06	0.01	0.00
Northern Baltic	0.39	0.37	0.18	0.05	0.01
Vahemadal	0.78	0.20	0.02	0.00	0.00

~~**Figure 3 Observation series of SWH used in the study.**~~

4 Verification methodology

In this section, a short overview of the verification procedure will be given. For background and more details regarding the verification measures, we refer to (Jolliffe and Stephenson, 2003)

For each measurement series of SWH, the corresponding forecast series for all forecast classes and for forecast range zero to 48 h for the grid point nearest to the position of the ~~station~~ wave measurement site was extracted from the model output.

For the deterministic and continuous forecast classes (LOW, LOWENSMEAN and HIGH), we use the conventional performance measures *root mean square error* (RMSE), defined as the square root of the time average of the sum of squared differences between forecast and observation:

$$RMSE(\tau) = \langle \left(h_{s,fcst}^{\tau} - h_{s,obs} \right)^2 \rangle$$

the bias

$$BIAS(\tau) = \langle h_{s,fcst}^{\tau} - h_{s,obs} \rangle,$$

and the correlation coefficient

$$CC = \frac{\langle (h_{s,fcst}^{\tau} - \langle h_{s,fcst}^{\tau} \rangle)(h_{s,obs} - \langle h_{s,obs} \rangle) \rangle}{\sqrt{\langle (h_{s,fcst}^{\tau} - \langle h_{s,fcst}^{\tau} \rangle)^2 \rangle \langle (h_{s,obs} - \langle h_{s,obs} \rangle)^2 \rangle}}$$

where $h_{s,obs}$ is the observed SWH and $h_{s,fcst}^{\tau}$ is a corresponding forecast with forecast range τ .

The RMSE is a positive definite quantitive measure, and smaller values mean a better forecast. The bias can take positive and negative values, and a good forecast has a numerically small value. The averaging, indicated by $\langle \cdot \rangle$, can be found based on all available values during the three-year period. Also, the RMSE and BIAS as function of $h_{s,obs}$ will be considered.

A framework for verifying probabilistic forecasts is the *continuous ranked probability score* (CRPS), defined as

$$CRPS(\tau) = \langle \int [F^{\tau}(h_s) - H(h_s - h_{s,obs})]^2 \, dh_s \rangle,$$

where $F^{\tau}(h_s)$ is the forecasted probability distribution, $h_{s,obs}$ is the observed value, and $H(\cdot)$ is the Heaviside step function. A small CRPS occurs when the median of the probabilistic forecasts are close to the observed values. Also a sharp probabilistic forecast with a small spread favors a small CRPS. This means that the best forecast is achieved when CRPS is small. CRPS can be applied to both the probabilistic forecast class LOWENS, as well as the deterministic forecast classes, LOW, LOWENSMEAN and HIGH, since these can be regarded as probabilistic forecasts with a step probability distribution. For the deterministic forecast classes, the CPRS equals the *mean absolute error*.

Besides the continuous and probabilistic forecasts, also the binary forecast of the SWH exceeding a specified threshold is considered. The performance measure used is the Brier Score, defined as

$$BS(\tau) = \langle (p - x)^2 \rangle,$$

where p is the forecasted probability with forecast range τ of exceeding the threshold and x takes the value of 1 or 0 dependent on whether the threshold actually was exceeded or not. The Brier Score is thus a positively definite measure, where values are between zero and one, and the lower the value, the better the forecast.

4.1 Calculation of confidence bands

All the measures described above are subject to sampling uncertainty; if they had been calculated on data from another time period than 2015-2017, they would have had different values. To estimate this sampling uncertainty and thereby obtain confidence bands, we applied a block bootstrapping procedure, where a large number of resampled series with the same length as the original series (three years) were created. A blocking length of one month was chosen. This choice takes the atmospheric decorrelation time scale of a few weeks into account and it allows a large number of different resampled series to be made.

Each resampled series is constructed as follows: The resampled series will contain three January's, and each of these is randomly chosen, with replacement, of the three January's from the original series. A similar procedure applies for February, etc. In this way, the resampled series are most likely different but the annual cycle is preserved. Both the observed series and the forecast series are resampled. For each pair of resampled series bootstrapped value of the performance measures are calculated. Repeating the resampling procedure, we obtain 1000 resampled values of the measures, from which their approximate statistical distribution and confidence bands can be calculated. As a standard, confidence bands (5/95%) are calculated by the bootstrap procedure described above and this allows for a quantitative inter-comparison of the performance measures for the different forecast classes: if the confidence bands do not overlap then there is a significance difference.

5 Verification of the wind forecasts

The two configurations of DMI-HIRLAM used (see Table 2) have been verified against available wind observations from Danish coastal stations, i.e. covering the western part, of the Baltic Sea, for the period 1 January 2015– 31 December 2017. For the S05 configuration, the ensemble mean is verified.

Table 5 Verification results for DMI-HIRLAM against Danish coastal stations for the period 1 January 2015– 31 December 2017. Positions of stations are marked on Figure 3 .

FCST	BIAS [ms^{-1}]		RMSE [ms^{-1}]		CC		Hit rate, error ≤ 2 ms^{-1}	
RANGE	S05(EM)	S03	S05(EM)	S03	S05(EM)	S03	S05(EM)	S03
Gedser (WMO 06149):								
0	0.48	0.46	1.57	1.56	0.90	0.91	0.82	0.82
6	0.43	0.46	1.58	1.62	0.90	0.90	0.81	0.80
12	0.45	0.49	1.66	1.72	0.89	0.89	0.79	0.78
18	0.45	0.49	1.73	1.83	0.88	0.87	0.77	0.76
24	0.45	0.51	1.77	1.91	0.87	0.86	0.76	0.74
36	0.42	0.51	1.92	2.03	0.85	0.84	0.73	0.70
48	0.44	0.46	2.03	2.16	0.82	0.81	0.70	0.67
Hammer Odde Lighthouse (WMO 06197):								
0	0.31	0.19	1.24	1.24	0.92	0.91	0.90	0.90
6	0.22	0.12	1.28	1.33	0.91	0.90	0.88	0.87
12	0.24	0.13	1.34	1.42	0.90	0.88	0.87	0.85
18	0.25	0.15	1.38	1.48	0.89	0.87	0.86	0.84
24	0.26	0.14	1.43	1.57	0.88	0.86	0.84	0.82
36	0.24	0.11	1.53	1.67	0.86	0.84	0.82	0.79
48	0.23	0.10	1.62	1.80	0.85	0.81	0.80	0.77

Table 5 summarizes verification results for 10m wind forecasts for the 3km-resolution S03 model and the ensemble mean of the 5km-resolution S05 model for two Danish coastal stations in the western part of the Baltic Sea. A comparison between the two model forecasts shows a small positive bias and RMS errors increasing with forecast range from approx. 1 ms^{-1} to approximately 2 ms^{-1} for 48h forecasts. The error of the ensemble mean forecasts generally increases less with forecast range than the error of the high-resolution forecasts. Similarly, the correlation and the hit rate (error ≤ 2 ms^{-1}) decrease with forecast range, but less so for the ensemble mean forecasts. That is, in terms of wind forcing the ensemble mean of the S05 model provides slightly more accurate forecasts than the higher resolution, deterministic S03 model, especially for the longer forecast ranges.

~~5~~6 Verification of forecasted SWH against observations

~~5.1~~6.1 Deterministic measures

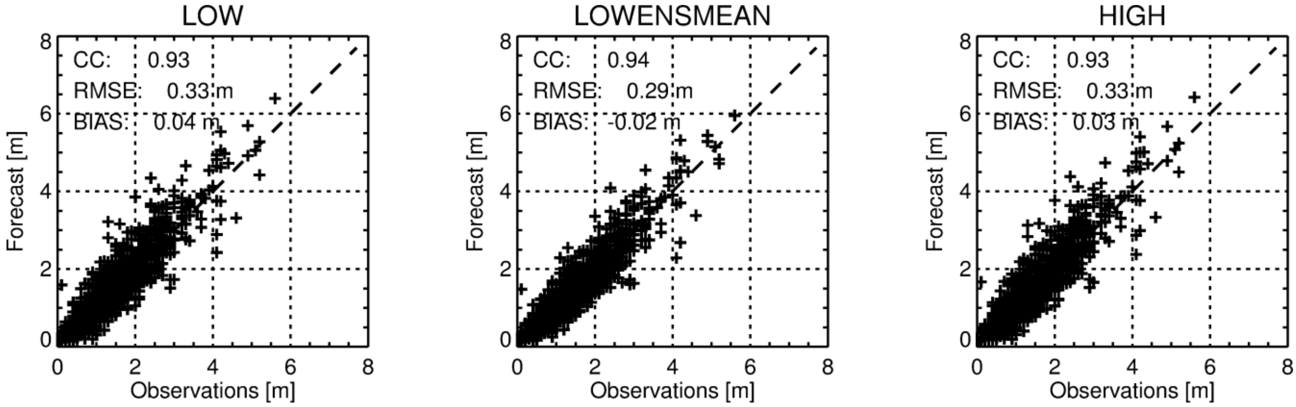


Figure 4~~44~~ Scatter plot of 24 h forecasts and corresponding observations of significant wave height at ~~station~~site Bothnian Sea for the LOW, LOWENSMEAN and HIGH forecast classes. Dotted line is the diagonal, representing a 1:1 agreement between observations and model.

To get an idea of the overall quality of the forecasts, Figure 4~~Figure 4~~ shows scatter plots between 24 h forecasted and observed SWH for station Bothnian Sea. The points are distributed along the diagonal in all three configurations with correlation coefficients above 0.9. The RMSE is 0.33 m for both LOW and HIGH but is lower at 0.29 m for the LOWENSMEAN forecasts, which also have the numerically lowest bias. Also for other ~~station~~sites, such as Arkona WR (see Figure 5~~Figure 5~~), the RMSE for LOWENSMEAN forecasts is lower than for the LOW and HIGH forecasts, and similarly for the bias. However, the scatter plot appears differently for this station, because there is a tendency for over-predicting high waves for all three forecast classes.

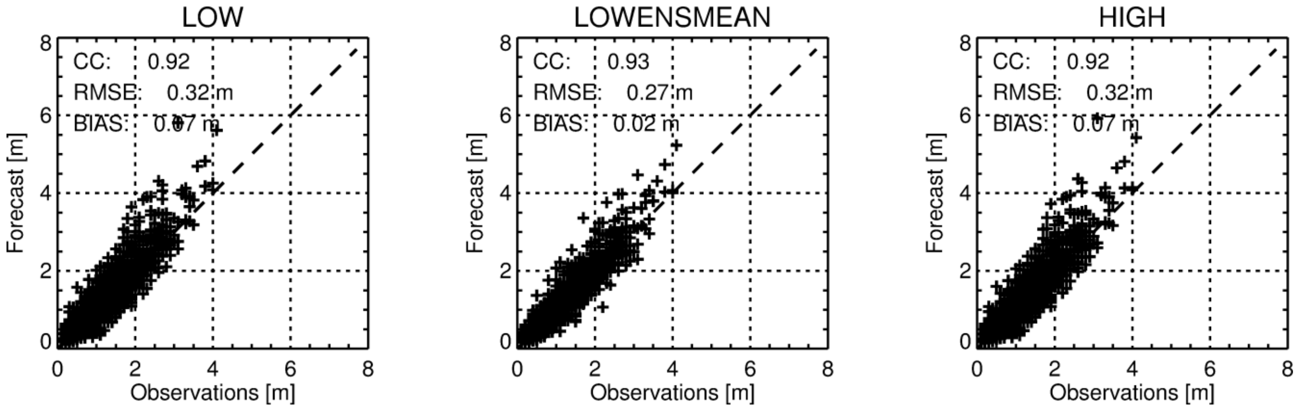


Figure 5~~5~~ As Figure 4~~Figure 4~~ but for ~~station~~site Arkona WR.

We now turn to the RMSE as function of forecast range, of which plots for all ~~station~~sites can be found in Figure ~~S1~~S2. For all sites, the RMSE increases slightly as function of forecast range. All sites except Vahemadal exhibit ~~Q~~qualitatively ~~the same picture is found for most other station~~similar behaviour: the RMSE ~~and BIAS for~~ for the LOW and HIGH forecasts are almost similar, while ~~they are~~it is lower for the LOWENSMEAN forecasts. ~~Plots for stations with qualitatively different behavior in this respect are shown in Figure 6.~~ Thus, ~~For~~ for Arkona WR (shown in Figure 6), Bothnian Sea and Darss Sill WR~~, all three forecasts have a non-zero RMSE for forecast range zero (the analysis). The reason for this is that this 'analysis' is a forecast with forecast range 6 h, made six hours before. The RMSE increases slightly as function of forecast range. T~~, the RMSE of the LOW and the HIGH forecasts ~~coincide to a large degree~~have overlapping confidence bands~~,~~. T~~while t~~he RMSE for LOWENSMEAN gradually diverges to a lower value (around ~~of~~ 5 cm) ~~lower~~ and for large forecast ranges ~~larger than 24 h~~, the confidence bands do not overlap with those for the LOW and HIGH forecast classes. The remaining sites except Vahemadal behave similarily, but with overlapping confidence bands even for the largest forecast ranges.

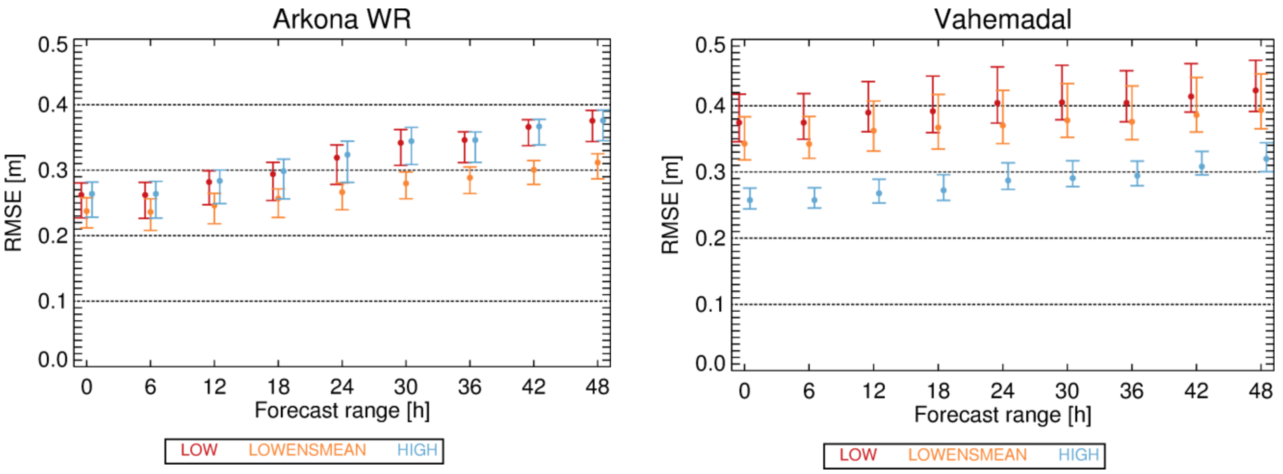


Figure ~~6~~6 RMSE for selected forecast ranges for Arkona WR (left panel) and Vahemadal (right panel) for LOW, LOWENSMEAN and HIGH forecasts. Error bars show 5/95% confidence bands calculated by bootstrapping.

~~Qualitatively the same picture is found for most other station: RMSE and BIAS for LOW and HIGH forecasts are almost similar, while they are lower for LOWENSMEAN forecasts. However, the RMSE values of the LOWENSMEAN forecasts are not for all stations well separated by non-overlapping confidence bands from RMSE of the other forecast classes.~~

The ~~station~~site Vahemadal (Figure 6) has a different behavior. For this ~~station~~site, the HIGH forecast class has a significantly smaller RMSE and with non-overlapping confidence bands with the RMSE of ~~than~~ the LOW and LOWENSMEAN forecasts~~, which have overlapping confidence bands~~. This ~~station~~site also has a non-negligible bias of around 12 cm for the HIGH and around 20 cm for the LOW and LOWENSMEAN forecasts; this bias is independent of forecast range (not shown).

~~5.1.1~~6.1.1 Performance depending on observed SWH

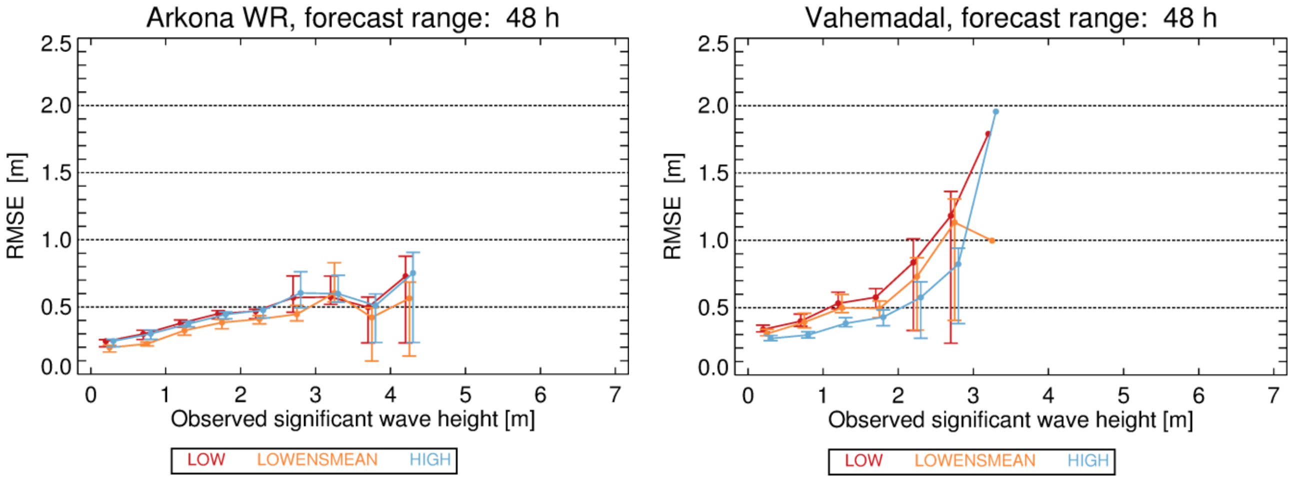


Figure 7~~7~~ RMSE as function of SWH for Arkona WR (left panel) and Vahemadal (right panel) for LOW, LOWENSMEAN and HIGH forecasts and forecast range 48 h. Error bars show 5/95% confidence bands calculated by bootstrapping.

The RMSE of the forecasts depends on the magnitude of the SWH. Plots for all ~~station~~sites for 24 and 48 h forecast range of RMSE as function of the SWH can be found in Figures ~~S2~~ S3 and ~~S3~~S4. The RMSE for Arkona WR and Vahemadal as function of the SWH for forecast range 48 h is shown in Figure 7~~Figure 7~~. The RMSE increases as function of the observed SWH for both ~~station~~sites. For Arkona WR, the LOWENSMEAN forecast class has the lowest RMSE, although with confidence bands overlapping with the other forecast classes. This behavior is seen ~~in~~ at all ~~station~~sites, except Vahemadal. For Vahemadal, the HIGH forecast class has the lowest RMSE, and up to a SWH of 2 m, the confidence band is well separated from the confidence bands of the other forecast classes.

Also the bias depends on the SWH. Plots for all ~~station~~sites for 24 and 48 h forecast range of ~~BIAS~~ the bias as function of the SWH are displayed in Figures ~~S4~~ S5 and ~~S5~~S6. For small SWH, the ~~BIAS~~ bias is close to zero for most ~~station~~sites. For some ~~station~~sites, the bias remains close to zero for increasing SWH, as shown for Arkona WR in left panel of Figure 8~~Figure 8~~, while for others it becomes different from zero for large values of SWH. The is no noticeable different in the bias of the different forecast classes, except for Vahemadal, shown in right panel of Figure 8~~Figure 8~~, where the HIGH forecast class has a significantly smaller bias than the other forecast classes.

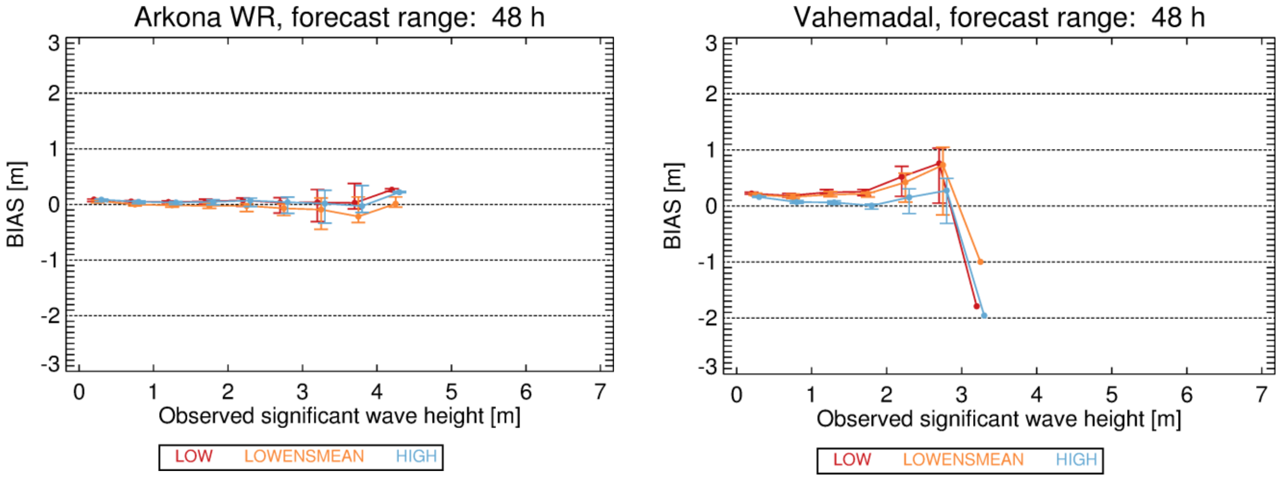


Figure 8~~8~~ ~~BIAS~~ Bias as function of SWH for Arkona WR (left panel) and Vahemadal (right panel) for LOW, LOWENSMEAN and HIGH forecasts and forecast range 24 h. Error bars show 5/95% confidence bands calculated by bootstrapping.

6.1.2 Forecasts during ‘Toini’ storm

The Toini storm on 11. January 2017, where a SWH of almost 8 m was recorded on Northern Baltic (Björkqvist et al., 2017a), is within our verification period.

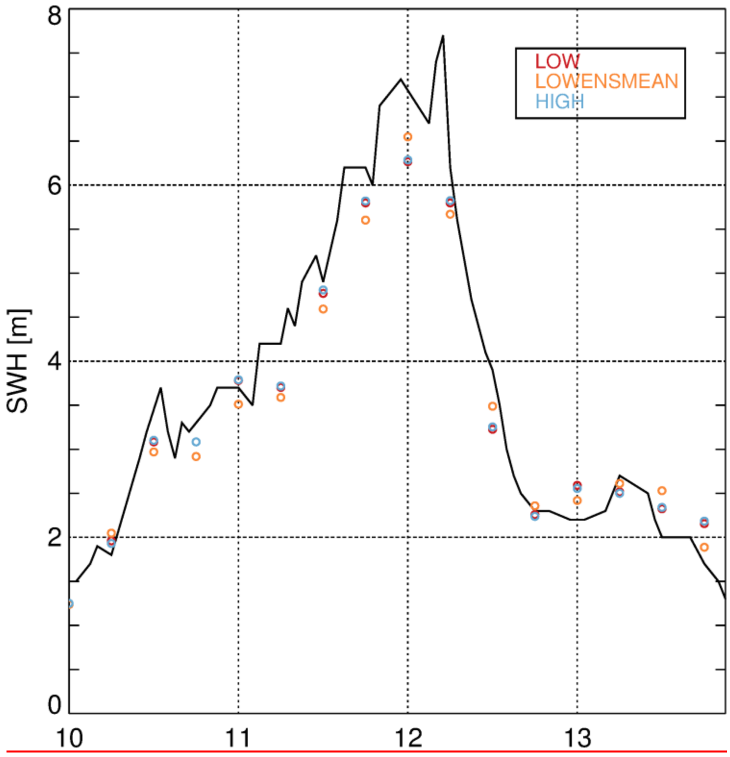


Figure 9 Observed SWH for Northern Baltic during, 10-13 January 2017, including the Toini storm. Open circles are 48 h forecasts.

Figure 9 shows the observed SWH at Northern Baltic during 10-13 January 2017, i.e. including the Toini storm, peaking in the early hours of 12 January, together with 48 h forecasts. In this case there is no apparent ‘best’ forecast. Near the peak, LOWENSMEAN performs best, but both before and after, the HIGH/LOW performs better. Further, that in most cases, the LOW and HIGH forecasts are very similar, indicating that the higher resolution does not improve the forecasts.

~~5.2~~6.2 Probabilistic metrics

The 11 ensemble members of the LOWENS forecast class defines a statistical distribution function, which is a probabilistic forecast of the wave conditions. Besides, the deterministic forecast classes LOW, LOWENSMEAN and HIGH may be regarded as probabilistic forecasts with probability one for the deterministically forecasted future state and probability zero for all other states.

As described in Section 4, we use CRPS to describe performance of probabilistic forecasts. CRPS for all ~~station~~sites for selected forecast ranges can be found in Figure S7~~6~~. As typical examples, Figure 10~~Figure 9~~ displays this plot for Arkona WR and Vahemadal.

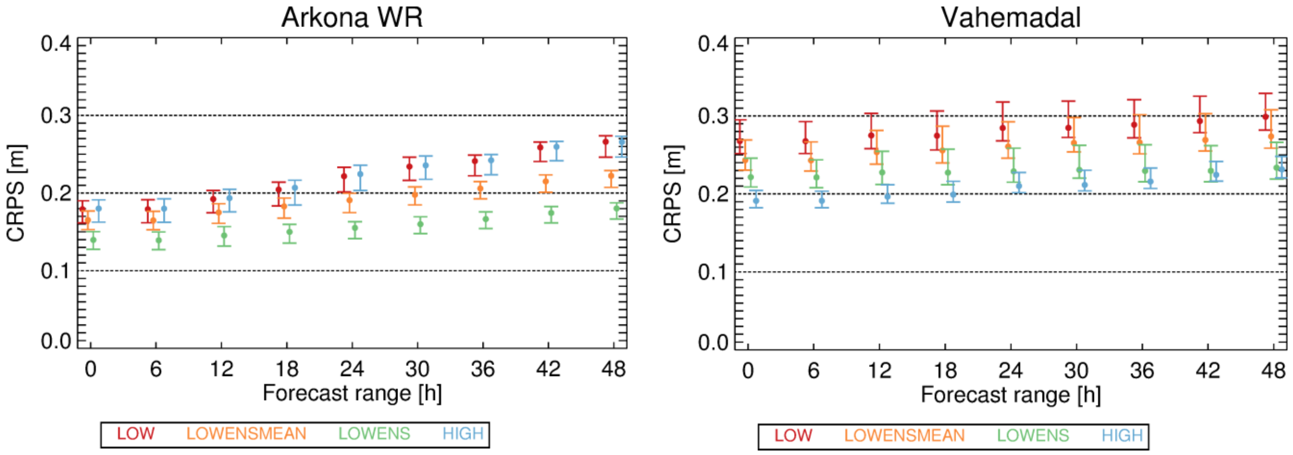


Figure 10~~9~~ CRPS for selected forecast ranges for Arkona WR (left panel) and Vahemadal (right panel)~~ WR~~ for LOW, LOWENSMEAN, LOWENS and HIGH forecasts. Error bars show 5/95% confidence bands calculated by bootstrapping.

~~This plot reveals that for~~All sites except Vahemadal behave qualitatively as Arkona WR~~,~~: the LOWENSMEAN forecast class has a~~ significantly~~ lower CRPS compared to both the HIGH and LOW classes, although the difference is significant (non-overlapping confidence bands) for Arkona WR, Bothnian Sea and Darsser Sill WR only, and only for the largest forecast ranges~~. This is most prominent for the large forecast ranges, where its confidence band is non-overlapping with the confidence band for other forecast classes~~. Furthermore, for all these sites, the LOWENS forecast class has an even lower CRPS, with confidence bands separated from those of all other forecasts classes.~~ This behavior is common among almost all stations.~~ Again, Vahemadal ~~is the exception~~behaves differently~~,~~; ~~where~~ here the HIGH forecast class has the best performance in terms of CRPS. However, for large forecast ranges, the LOWENS forecast class tends to perform equally well.

~~5.3~~6.3 Binary forecasts

For the probabilistic LOWENS forecast class, a binary forecast can be derived as the probability of exceeding a defined threshold of SWH. For the deterministic forecast classes: LOW, LOWENSMEAN and HIGH, this probability of exceedance is either zero or one. As described in Section 4, the Brier Score is used as performance measure for probabilistic, binary forecasts.

The Brier Score as function of threshold is shown for all ~~station~~sites in Figures S7 and S8. Figure 11~~Figure 10~~ shows the Brier Score as function of threshold for Arkona WR and Vahemadal for 48 h forecast range. For Arkona WR, the Brier Score for the LOWENS forecast class is the smallest, however the confidence intervals overlap with confidence intervals from the other forecasts above 2 m threshold. Also the LOWENSMEAN forecast class has low Brier Score. This behavior is common to~~ almost~~ all ~~station~~sites except Vahemadal. ~~The exception is again~~For Vahemadal, ~~where~~ the Brier Score is smallest for the HIGH forecasts for thresholds above 1 m.

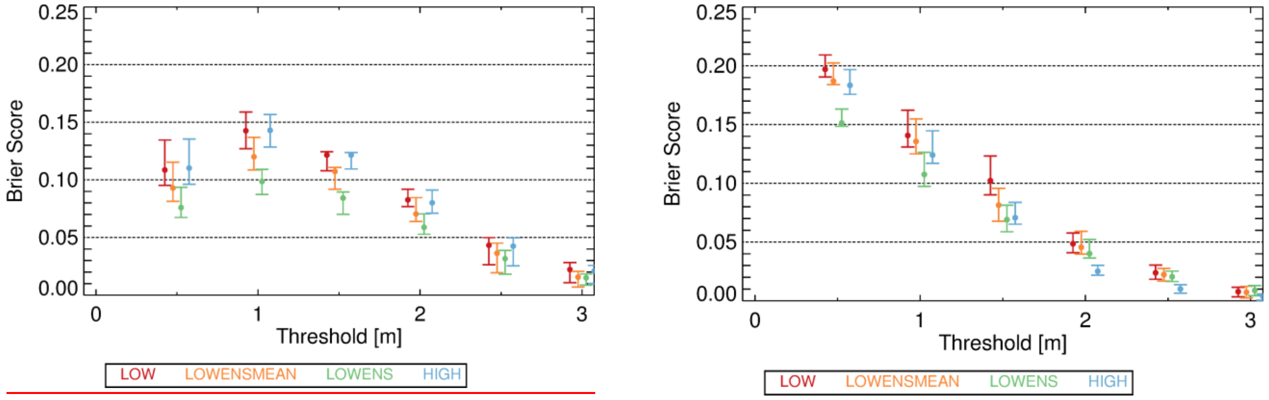


Figure 11~~10~~ Brier score for Arkona WR (left panel) and Vahemadal (right panel) for binary forecast for forecast range 48 h.

6.4 Rank histogram

Rank histograms serve the purpose of illustrating the reliability of probabilistic ensemble forecasts. It is a histogram of the rank of the observation, when the observation and all ensemble members of the corresponding forecast are pooled together. If the observations and the ensemble members belong to the same distribution, then the rank histogram will be flat, while a U-shaped histogram indicates too small variance within the ensemble members. For more discussion, see Jolliffe and Stephenson (2003).

Rank histograms for all wave measurement sites for forecast range 24 and 48 h are shown in Figure S10 and S11 for forecast range 24 resp. 48 h. We note that all histograms show the U-shape, indicating an unrealistically small variance within the ensembles. For most sites the U-shape is symmetric, except for Vahemadal, where the U-shape is strongly asymmetrical. This corresponds well with the bias mentioned in Section 6.1.

~~6~~7 Discussion

Our main finding in the previous section is that for most ~~stations~~wave measurement sites included in this study, the LOWENSMEAN and the LOWENS forecast classes have a performance superior to the LOW and HIGH forecast classes. Only for one ~~station~~site results are different; namely that the HIGH forecast class has the superior performance. These conclusion hold, whether based on overall RMSE, CRPS or the Brier score.

~~6.1~~7.1 Comparison with other operational forecast~~s~~ systems

Multi-year verification results from two operational deterministic wave forecast systems have been published, and can be compared to results from the present study. Both these systems are based on the third generation WAM; the system described in (Tuomi et al., 2008) has about 22 km horizontal resolution, while the system described in (Tuomi et al., 2017) has 1 naut. mile horizontal resolution.

For certain ~~station~~sites, the RMSE of the 6 hour forecasts of SWH are available for at least one of the aforementioned forecast systems in addition to the DMI-WAM forecasts; thus comparison of the systems is possible. All ~~station~~sites have a water depth of more than 46 m and therefore represent offshore conditions.

Table 6 Comparison of RMSE for SWH of 6h forecast runs for selected ~~station~~sites. FIMR values are from (Tuomi et al., 2008) and FMI values are from (Tuomi et al., 2017)

	FIMR	FMI	DMI LOW	DMI LOWENSMEAN	DMI HIGH
Horizontal resolution WAM	~ 22 km	1 naut. mile	10 km	10 km	5 km
Horizontal resolution NWP	~ 22 km	2.5 km	3 km	5 km	3 km
Arkona WR	-	0.28	0.26	0.24	0.26
Bothnian Sea	-	0.28	0.25	0.23	0.25
Finngrundet WR	-	0.27	0.24	0.22	0.23
Helsinki Buoy	0.25	0.26	-	-	-
Northern Baltic	0.31	0.26	0.24	0.23	0.24

From Table 6 ~~Table 4~~ one can see that for the ~~station~~sites considered, the LOWENSMEAN has the lowest RMSE. This supports the finding of this study that for offshore conditions, there is no reason to improve the resolution further than that of the LOW configuration. In addition, the results emphasize the value of describing the uncertainties of in the atmospheric forcing by introducing ensembles, as this leads to a lower RMSE of the forecasts. This is also in line with our findings in the previous section.

Test runs of a few months duration of deterministic and ensemble wave forecasts of SWH for the Baltic Sea (Behrens, 2015) also shows slight improvement of ensemble mean forecasts, compared to deterministic forecasts, and thus supports our findings.

Fore completeness, we remind the reader that the cases compared in Table 6 have different wind forcing and probably also different version of WAM. Therefore the differences seen cannot with certainty be attributed to differences in horizontal resolution.

7.2 Limitations of the study

7.2.1 Length of verification period

Operational centers typically renew their computer installations every 5-6 years with about an order of magnitude increase in performance. At DMI, a new installation was introduced primo 2016, allowing the HIGH and LOWENS configurations to replace the LOW configuration. Presently (medio 2018) the system is mid-term upgraded and this makes it appropriate to do the intercomparison now as a guidance for any changes in the operational setup.

Thus, the operational forecasts performed on the present system, supplemented by delayed-mode forecasts has determined the three-year verification period used in our study. A longer verification period could evidently have reduced the sampling uncertainty in the analyses and thereby sharpened the conclusions. On the other hand, the three-year verification is not short compared to other studies, e.g. Bunney and Saulter (2015) or Tuomi et al.(2017)

~~6.1.1~~7.2.2 Choice of observational base

The present verification is based on observation in near-hourly resolution from a number of ~~station~~sites in the Baltic Sea. Therefore, in the major parts of the Baltic Sea, verification is not possible, which puts a limit on how strong conclusions can be made.

SWH derived from satellite-borne altimeters (Kudryavtseva and Soomere, 2016) offers an alternative, which could be pursued in a future study. These data has a fair spatial data coverage but at the cost of a temporal resolution of one day or less. This means that maximum wave heights connected to severe storms may easily be missed. Nevertheless, these data has proven useful for verification in the Baltic Sea by (Tuomi et al., 2011)

~~6.2~~7.3 Effect of sea ice coverage

The main effect of sea ice on formation of waves is to limit the fetch. Furthermore, when a developed wave field approach an ice-covered area, the wind and the waves decouple, so that the waves act more like swell, propagating through ice-covered areas while losing energy by breaking up the ice cover. The WAM model does not account for such interactions, and sea ice, when dense enough, act as a solid shield that effectively remove all local wave energy in the model. It is implicitly assumed that dense ice will also be thick enough for this to approximately correct. In the Baltic Sea, that may not always be the case, and therefore sea ice occurrence may represent a systematic error source in the present study. Another effect of sea ice in the Baltic is that the wave observing systems are withdrawn, when ice is expected. This may cause a systematic bias in the verification analysis, if strong winds during winter are left out.

Based on Copernicus sea ice charts produced by the Finnish Meteorological institute the ice conditions for the Baltic have been evaluated. The Finnish ice charts are produced on a grid of approximately 1 km^2 with a temporal resolution of approximately one day in the ice season. Data is available from 2010 onwards. The average ice conditions for February for all years and the three years in focus can be found in Figure S12. All three years 2015-2017, and in particular 2015, have a smaller ice cover relative to the period 2010-2018.

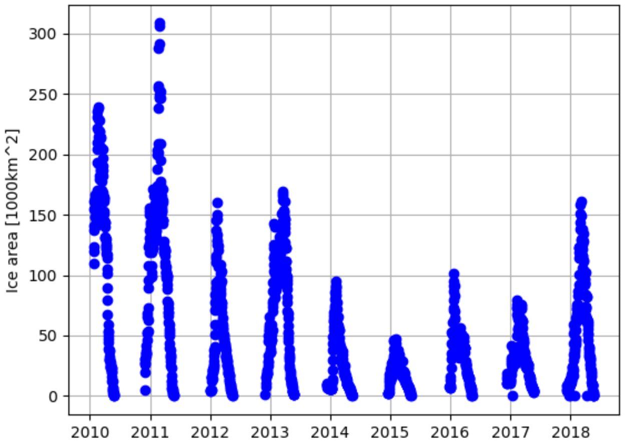


Figure 12 Integrated sea ice area of the Baltic Sea based on Finnish ice charts

Another way to illustrate this is considering the Baltic Sea integrated sea ice area, depicted in Figure 12, which shows that the years 2015-2017 have the lowest sea ice area over the whole period 2010-2018.

Therefore, we may anticipate that systematic errors arise from the occurrence of sea ice are relatively small.

~~7~~8 Conclusion

For most ~~station~~sites, we find that the HIGH forecast class does not perform superior to the LOW forecast class in forecasting SWH. These ~~station~~sites are all positioned well away from coasts in deep water and are thus freely exposed from all directions. This ~~indicates~~ suggests that the resolution of the bathymetry and the spectral resolution are adequate. For these offshore ~~station~~sites, introducing ensembles ~~improves~~ increases the performance of the forecasts, whether as in the LOWENSMEAN deterministic forecasts or in the LOWENS probabilistic forecasts. A similar conclusion generally holds for the binary forecast of exceeding a threshold.

For one site, Vahemedal just outside Tallin, the HIGH forecast class performs better than the other classes. The bathymetry near Vahemedal is complex and relatively shallow, thus the bathymetry affects the wave field and an improved description will therefore improve the modeled wave field. ~~For one station, Vahemadal, the HIGH forecast class performs better than the other classes. Vahemadal is a coastal station, situated just outside Tallinn in a complicated bathymetry with an island nearby and therefore shielded from many directions~~. ~~This can explain that better description of bathymetry and better description of short waves improves the forecast.~~ Further verification with near-coast stations may reveal whether this conclusion holds in general for coastal areas.

For high wave heights, there are significant systematic biases for most ~~station~~sites shared among all three forecast configurations. These are most probably to be ascribed to model deficiencies and act to mask any differences in performance between the different forecast classes. Also the RMSE becomes large for large observed SWH. This is expected since small timing errors in the predicted wave time series will have larger impacts on the model-observation match-up when the SWH is large.

~~Based on the above, we hypothesize~~The present study therefore suggests that for offshore conditions, there are no indications of further increase of the resolution of the WAM model will result in enhanced forecast performance. In addition, the results show that introducing ensembles increases the performances. This is both true for deterministic forecast in the form of ensemble mean and for probabilistic forecast.

For nearshore conditions conclusions are based on only one ~~station~~site, but results from this indicates that increasing the resolution gives better forecasts, while introducing ensembles does not. This can be due to both enhanced spatial resolution, allowing a better representation of shadow and shallow water effects, and/or spectral resolution.

The results of the present study thus underpins that a wave model setup with an equidistant grid cannot deliver optimal wave forecasts for both coastal and offshore conditions. This is particularly true for the Baltic Sea, where very small spatial scales are found in the archipelago near the coasts of Sweden and Finland (Björkqvist et al., 2017b). Besides implementing a 0.1 naut. miles model, these authors improved forecasts by introducing semi-empirical modifications to the wave model. The issue is described in Cavaleri et al. (2018), where other approaches are discussed. These include one-way nesting, used in the present study (see Section 2), multi-cell grids (Bunney and Saulter, 2015), and triangular unstructured grids (e.g. Zijlema, 2010). These techniques may be worth testing for the Baltic Sea.

Data availability. Model data is available from the authors upon request, whereas wave observations can be found on the CMEMS server.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was carried out under the EfficienSea2 project and supported by European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 636329. Observational data were kindly provided by Copernicus Marine Environmental Monitoring System (CMEMS). We thank two anonymous referees for valuable suggestions.

References

Alari, V., Staneva, J., Breivik, Ø., Bidlot, J.-R., Mogensen, K. and Janssen, P.: Surface wave effects on water temperature in the Baltic Sea: simulations with the coupled NEMO-WAM model, Ocean Dyn., 66(8), 917–930, 2016.

Alves, J.-H. G., Wittmann, P., Sestak, M., Schauer, J., Stripling, S., Bernier, N. B., McLean, J., Chao, Y., Chawla, A., Tolman, H. and others: The NCEP–FNMOC combined wave ensemble product: Expanding benefits of interagency probabilistic forecasts to the oceanic environment, Bull. Am. Meteorol. Soc., 94(12), 1893–1905, 2013.

Amante, C. and Eakins, B. W.: ETOPO1 1 ARC-MINUTE GLOBAL RELIEF MODEL: PROCEDURES, DATA SOURCES AND ANALYSIS., 2009.

Battjes, J. A. and Janssen, J. P. F. M.: Energy Loss and Set-Up Due to Breaking of Random Waves, in Coastal Engineering 1978., 1978.

Behrens, A.: Development of an ensemble prediction system for ocean surface waves in a coastal area, Ocean Dyn., 65(4), 469–486, doi:10.1007/s10236-015-0825-y, 2015.

Björkqvist, J.-V., Tuomi, L., Tollman, N., Kangas, A., Pettersson, H., Marjamaa, R., Jokinen, H. and Fortelius, C.: Brief communication: Characteristic properties of extreme wave events observed in the northern Baltic Proper, Baltic Sea, Nat. Hazards Earth Syst. Sci., 17(9), 1653–1658, doi:10.5194/nhess-17-1653-2017, 2017a.

Björkqvist, J.-V., Tuomi, L., Fortelius, C., Pettersson, H., Tikka, K. and Kahma, K. K.: Improved estimates of nearshore wave conditions in the Gulf of Finland, J. Mar. Syst., 171, 43–53, doi:10.1016/j.jmarsys.2016.07.005, 2017b.

Bunney, C. and Saulter, A.: An ensemble forecast system for prediction of Atlantic–UK wind waves, Ocean Model., 96, 103–116, doi:10.1016/j.ocemod.2015.07.005, 2015.

Cao, D., Tolman, H. L., Chen, H. S., Chawla, A. and Gerald, V. M.: Performance of the ocean wave ensemble forecast system at NCEP, in The 11th International Workshop on Wave Hindcasting & Forecasting and 2nd Coastal Hazards Symposium. [online] Available from: http://nopp.ncep.noaa.gov/mmab/papers/tn279/mmab279.pdf (Accessed 22 September 2017), 2009.

Carrasco, A. and Saetra, Ø.: A limited-area wave ensemble prediction system for the Nordic Seas and the North Sea, Norwegian Meteorological Institute., 2008.

Cavaleri, L., Fox-Kemper, B. and Hemer, M.: Wind Waves in the Coupled Climate System, Bull. Am. Meteorol. Soc., 93(11), 1651–1661, doi:10.1175/BAMS-D-11-00170.1, 2012.

Cavaleri, L., Abdalla, S., Benetazzo, A., Bertotti, L., Bidlot, J.-R., Breivik, ø., Carniel, S., Jensen, R. E., Portilla-Yandun, J., Rogers, W. E., Roland, A., Sanchez-Arcilla, A., Smith, J. M., Staneva, J., Toledo, Y., van Vledder, G. P. and van der Westhuysen, A. J.: Wave modelling in coastal and inner seas, Prog. Oceanogr., doi:10.1016/j.pocean.2018.03.010, 2018.

Günther, H., Hasselmann, S. and Janssen, P. A. E. M.: The WAM Model cycle 4, World Data Center for Climate (WDCC) at DKRZ., 1992.

Hasselmann, K., Barnett, T., Bouws, E., Carlson, H., Cartwright, D., Enke, K., Ewing, J., Gienapp, H., Hasselmann, D., Kruseman, P., Meerburg, A., M{ü}ller, P., Olbers, D., Richter, K., Sell, W. and Walden, H.: Measurements of wind-wave growth and swell decay during the {Joint North Sea Wave Project}, Deut Hydrogr Z, 8(12), 1–95, 1973.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, John Wiley & Sons., 2003.

Kudryavtseva, N. A. and Soomere, T.: Validation of the multi-mission altimeter wave height data for the Baltic Sea region, Est. J. Earth Sci., 65(3), 161, doi:10.3176/earth.2016.13, 2016.

Saetra, Ø. and Bidlot, J.-R.: Assessment of the ECMWF Ensemble Prediction Sytem for Waves and Marine Winds, European Centre for Medium-Range Weather Forecasts., 2002.

Schaffer, J., Timmermann, R., Arndt, J. E., Kristensen, S. S., Mayer, C., Morlighem, M. and Steinhage, D.: A global, high-resolution data set of ice sheet topography, cavity geometry, and ocean bathymetry, Earth Syst. Sci. Data, 8(2), 543–557, doi:10.5194/essd-8-543-2016, 2016.

She, J., Allen, I., Buch, E., Crise, A., Johannessen, J. A., Le Traon, P.-Y., Lips, U., Nolan, G., Pinardi, N., Reißmann, J. H., Siddorn, J., Stanev, E. and Wehde, H.: Developing European operational oceanography for Blue Growth, climate change adaptation and mitigation, and ecosystem-based management, Ocean Sci., 12(4), 953–976, doi:10.5194/os-12-953-2016, 2016.

Tuomi, L., Kangas, A., Leinonen, J. and Boman, H.: The Accuracy of FIMR Wave Forecasts in 2002-2005., 2008.

Tuomi, L., Kahma, K. K. and Pettersson, H.: Wave hindcast statistics in the seasonally ice-covered Baltic sea., Boreal Environ. Res., 16, 2011.

Tuomi, L., Vähä-Piikkiö, O. and Alari, V.: Baltic Sea Wave Analysis and Forecasting Product BALTICSEA_ANALYSIS_FORECAST_WAV_003_010, 2017.

Zijlema, M.: Computation of wind-wave spectra in coastal waters with SWAN on unstructured grids, Coast. Eng., 57(3), 267–277, doi:10.1016/j.coastaleng.2009.10.011, 2010.

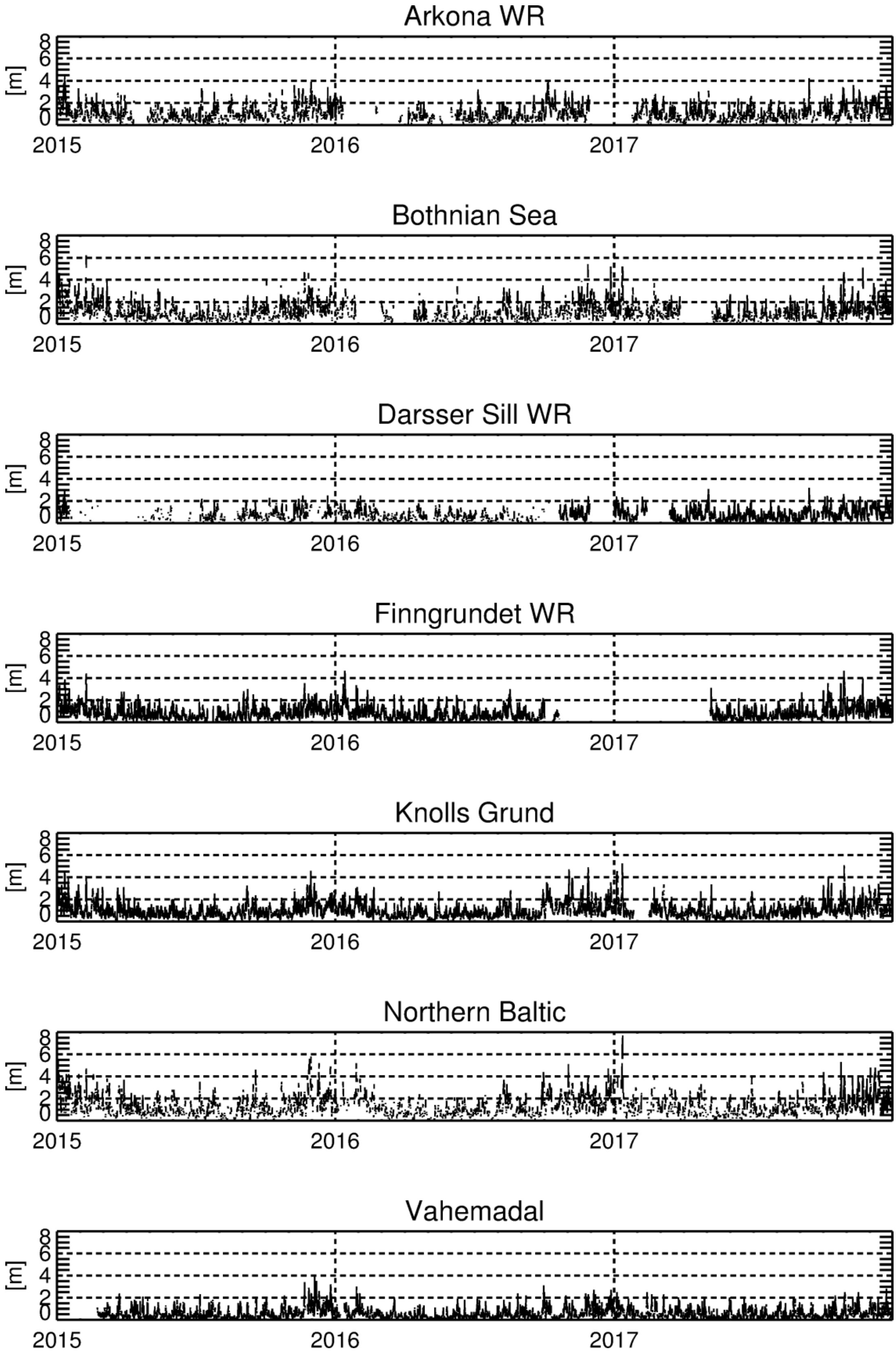


Figure S 1 Observation series of SWH used in the study.

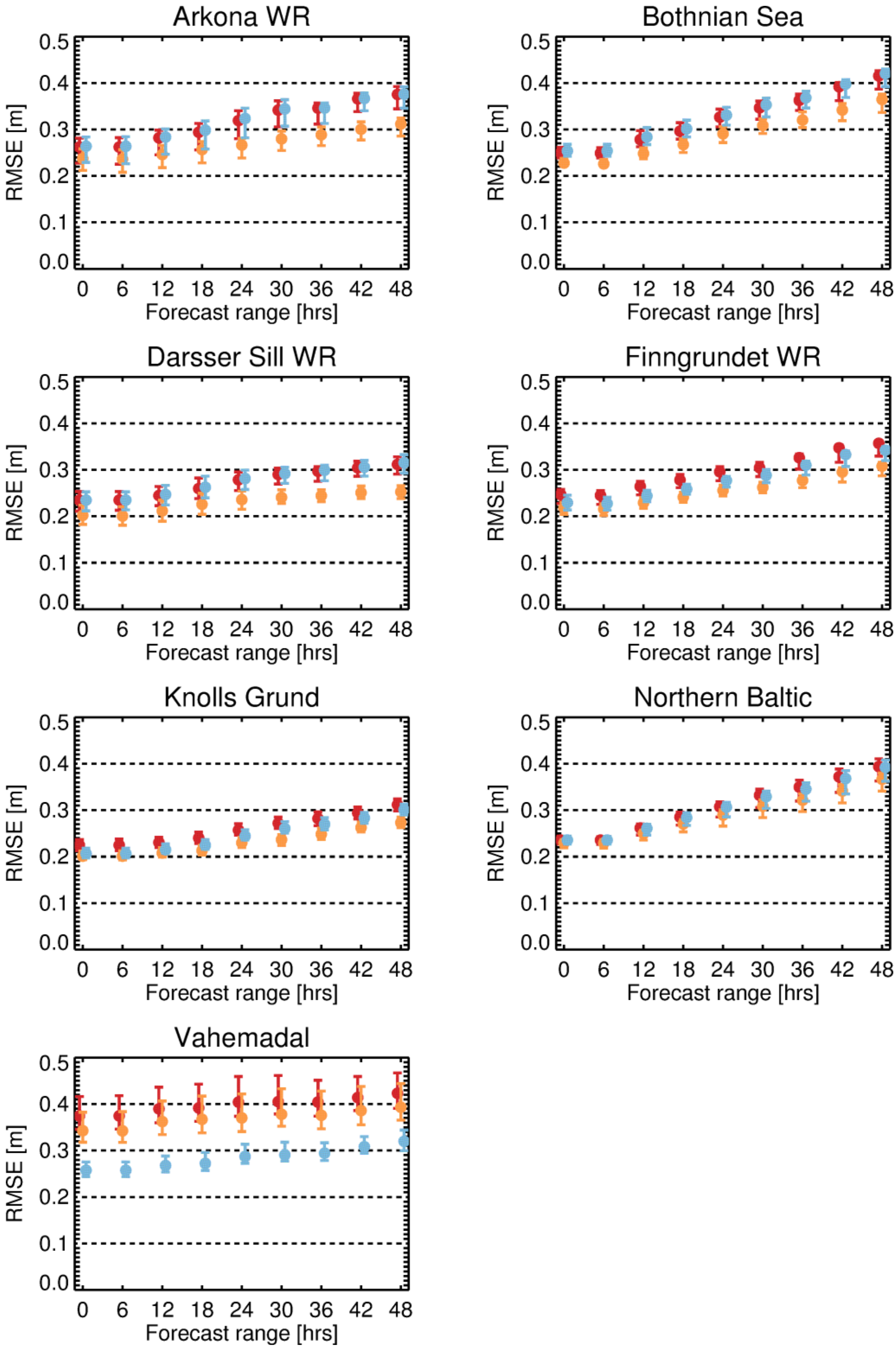


Figure S 2 RMSE for selected forecast ranges for all stations for LOW, LOWENSMEAN and HIGH forecasts. Error bars show 5/95% confidence bands calculated by bootstrapping

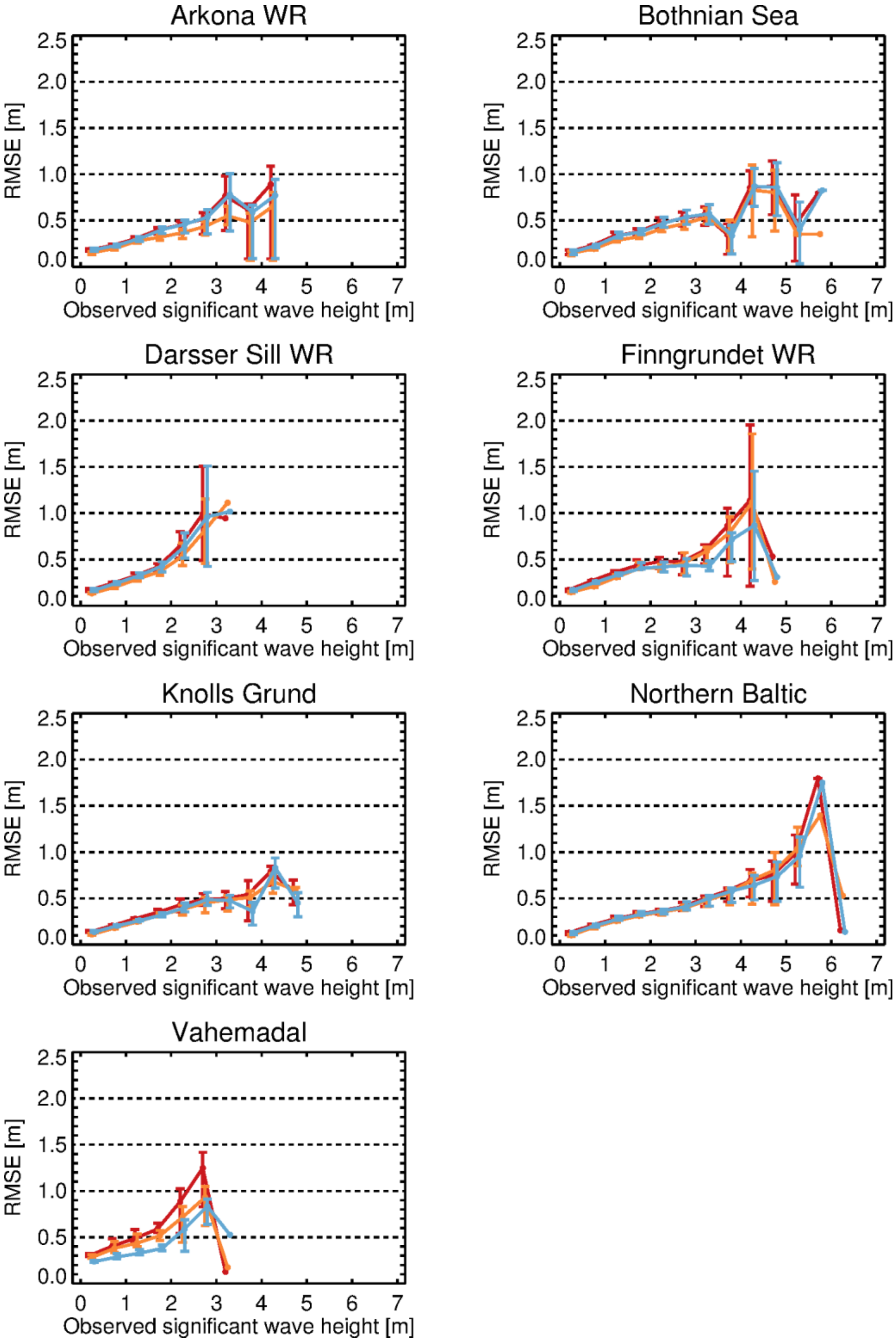


Figure S 3 RMSE as function of SWH for all stations for LOW, LOWENSMEAN and HIGH forecasts and forecast range 24 h. Error bars show 5/95% confidence bands calculated by bootstrapping. No confidence band is given when too few data points are available.

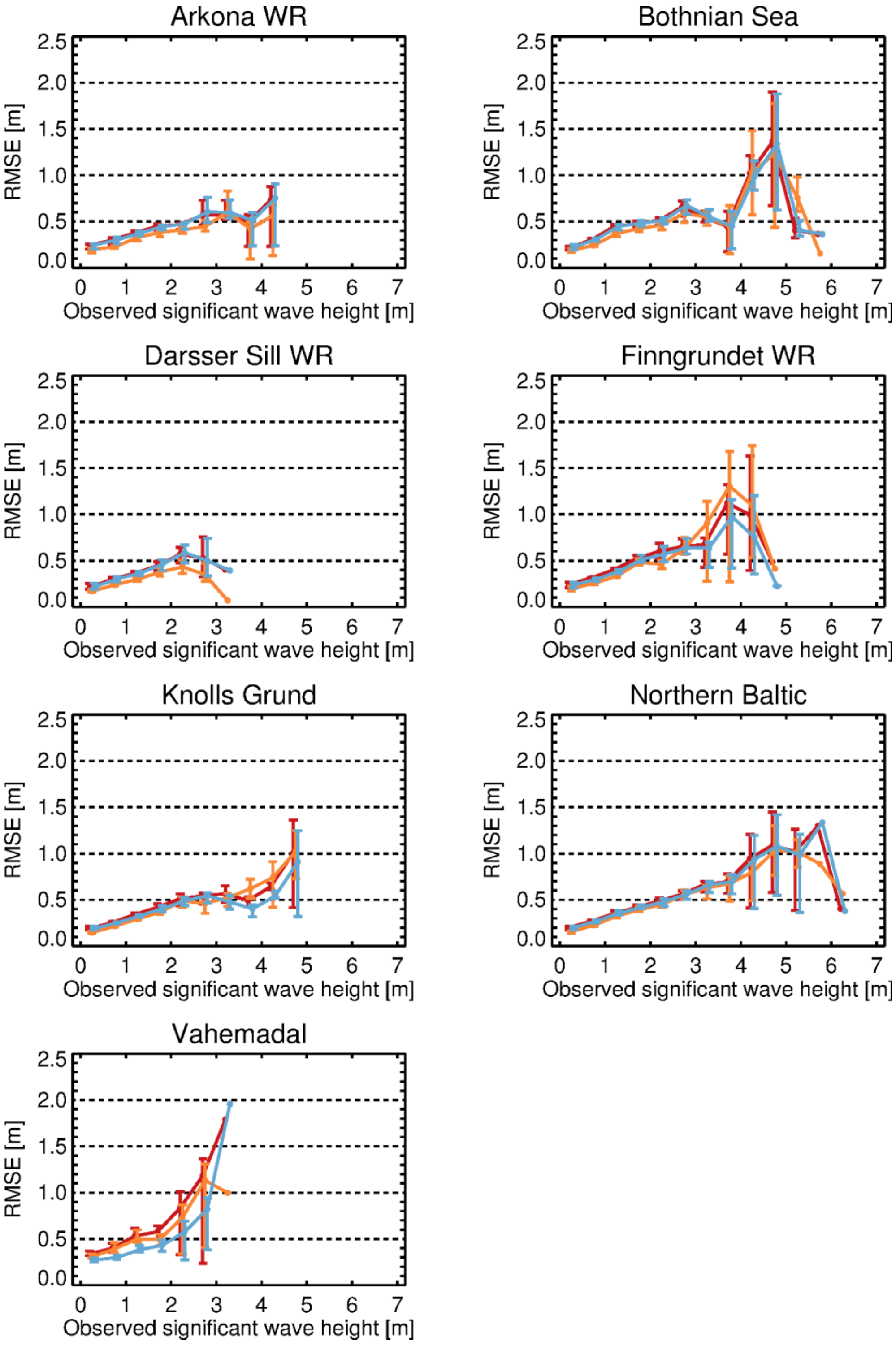


Figure S 4 As Figure S 3 but for 48 h forecast range.

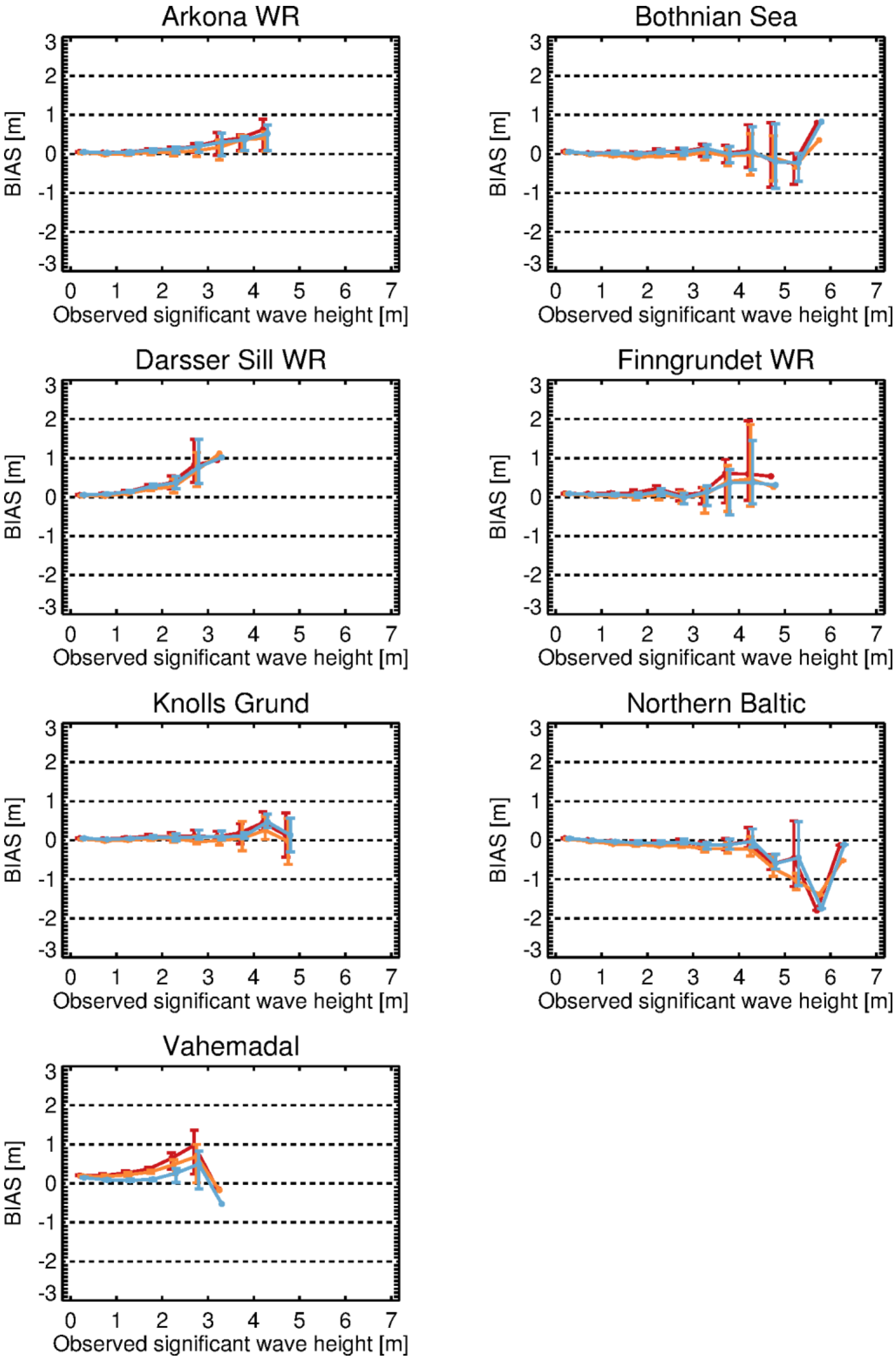


Figure S 5 Bias as function of SWH for all stations for LOW, LOWENSMEAN and HIGH forecasts and forecast range 24 h. Error bars show 5/95% confidence bands calculated by bootstrapping. No confidence band is given when too few data points are available.

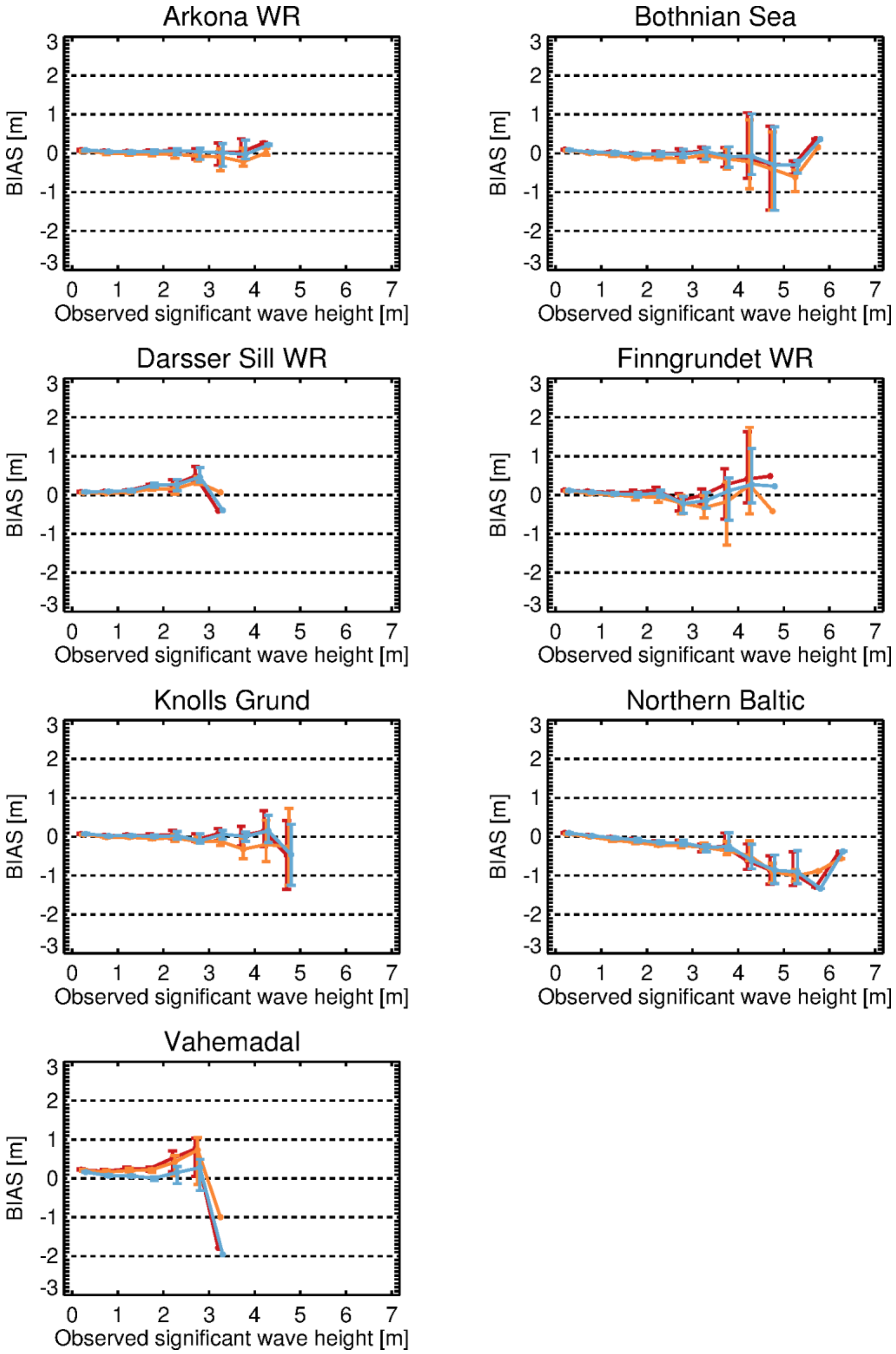


Figure S 6 As Figure S 5 but for forecast range 48 h.

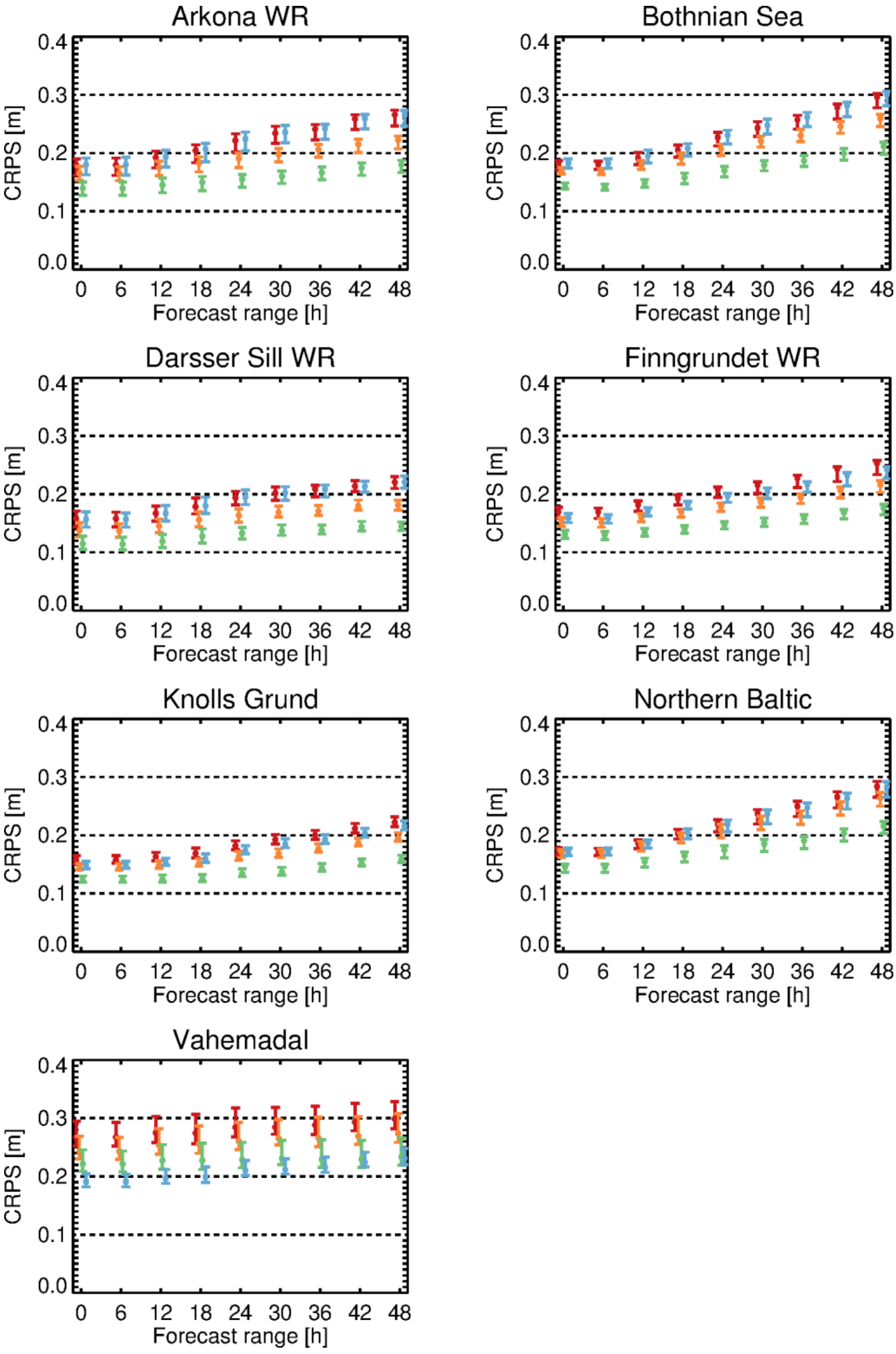


Figure S 7 CRPS for selected forecast ranges for all stations for LOW, LOWENSMEAN, LOW and HIGH forecasts. Error bars show 5/95% confidence bands calculated by bootstrapping.

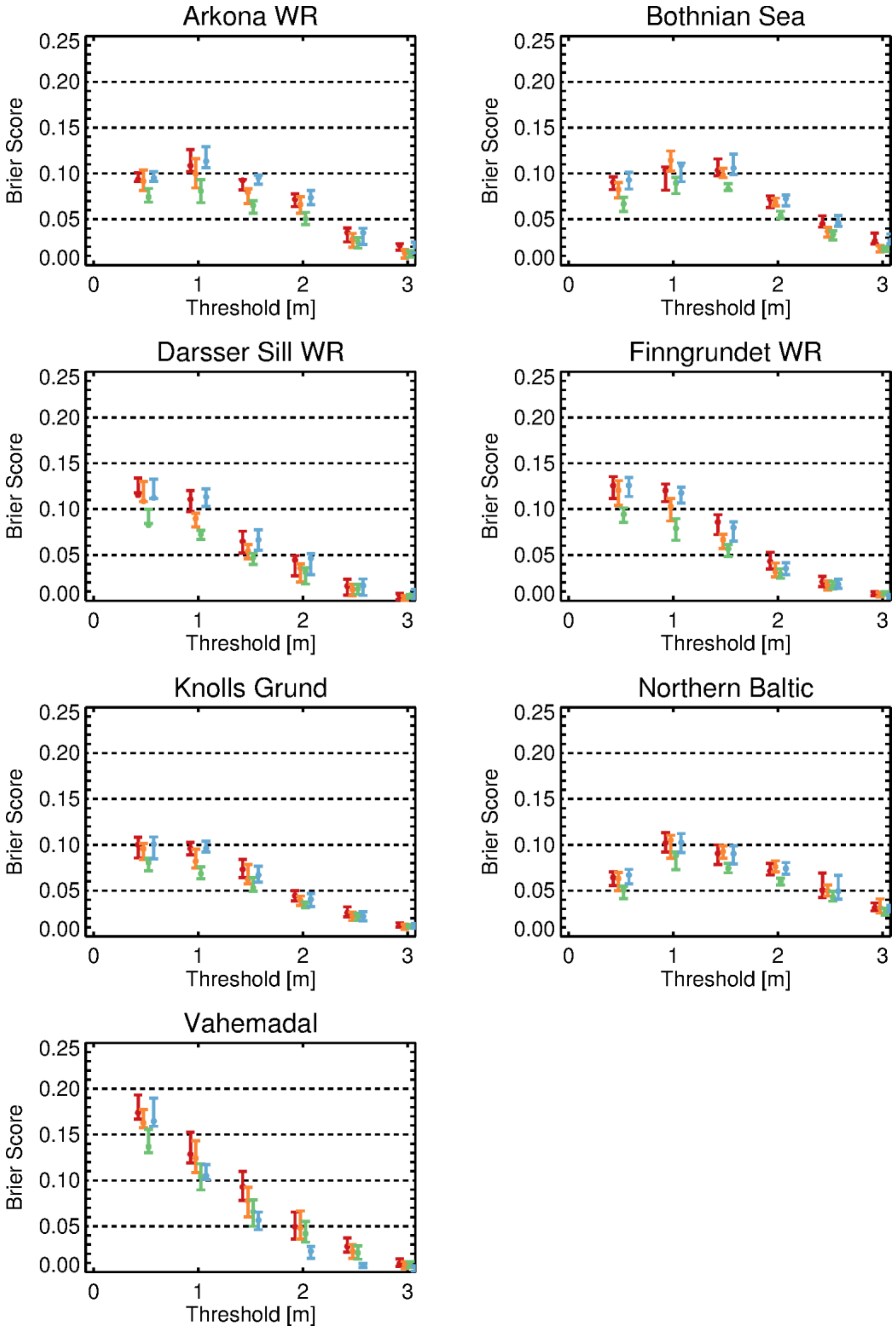


Figure S 8 Brier score for different thresholds for all stations for LOW, LOWENSMEAN, LOW and HIGH and forecast range 24 hrs.

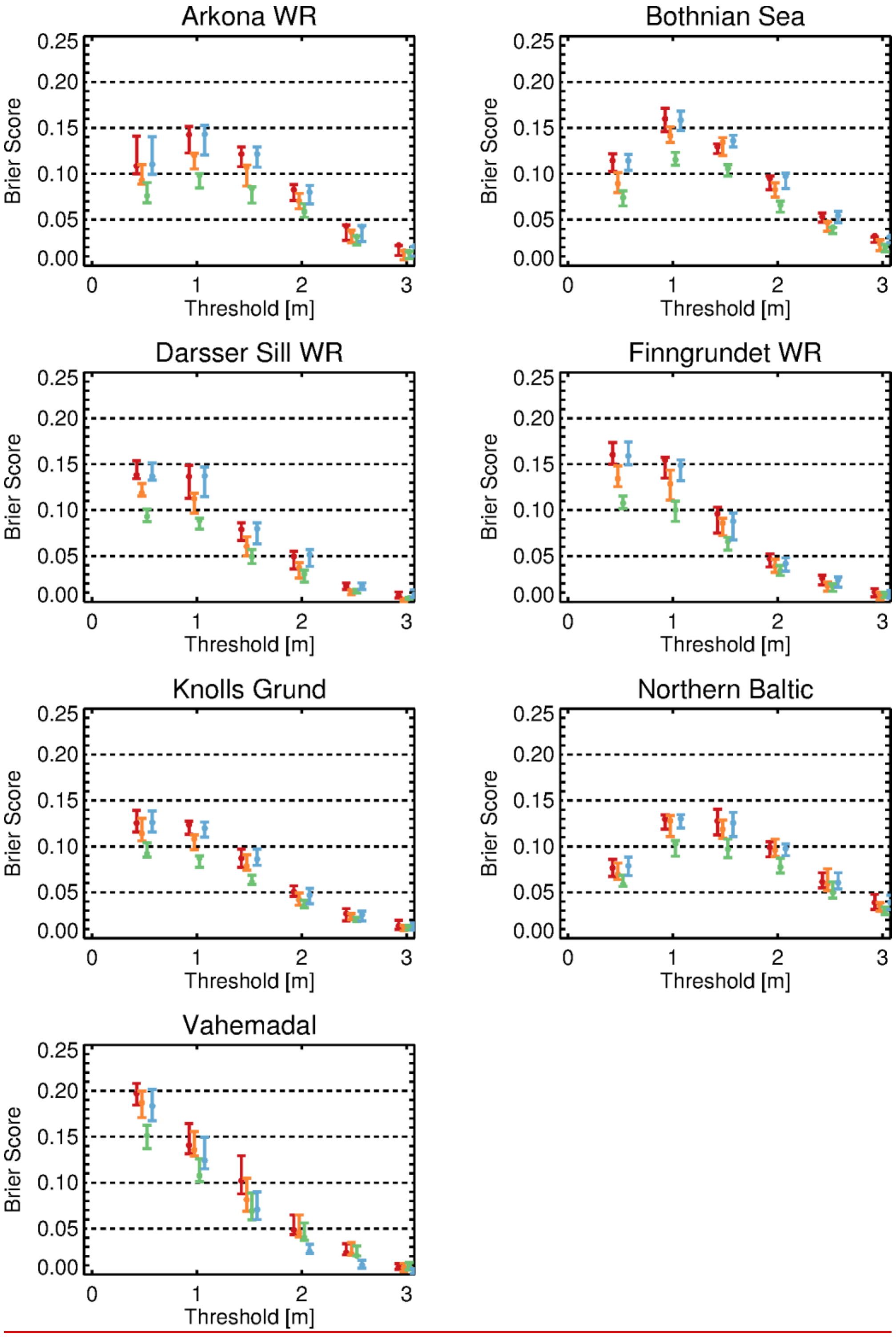


Figure S 9 As Figure S 8 but for forecast range 48 hrs.

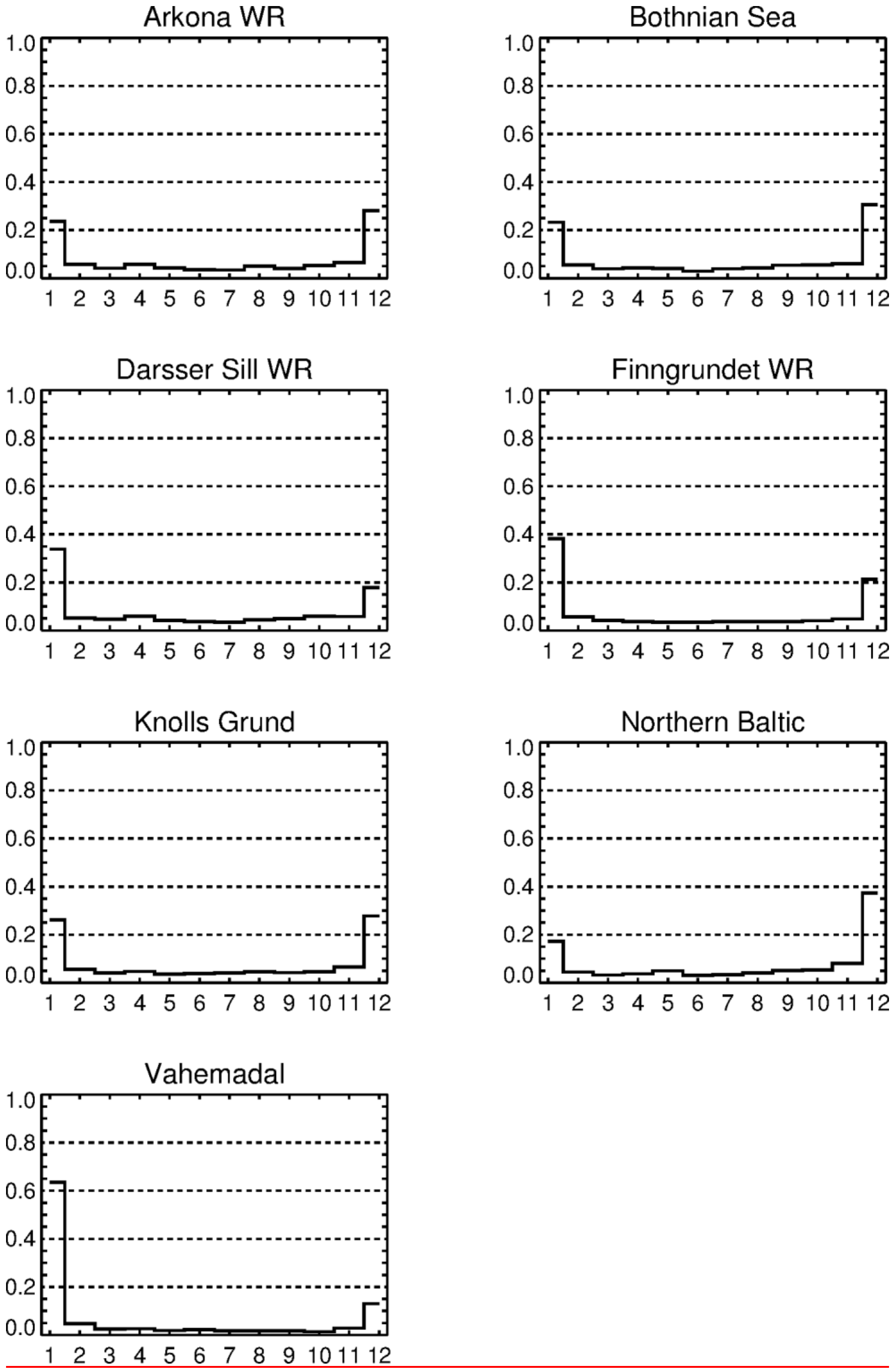


Figure S 10 Rank histograms for different thresholds for all stations for LOW, LOWENSMEAN, LOW and HIGH and forecast range 24 hrs.

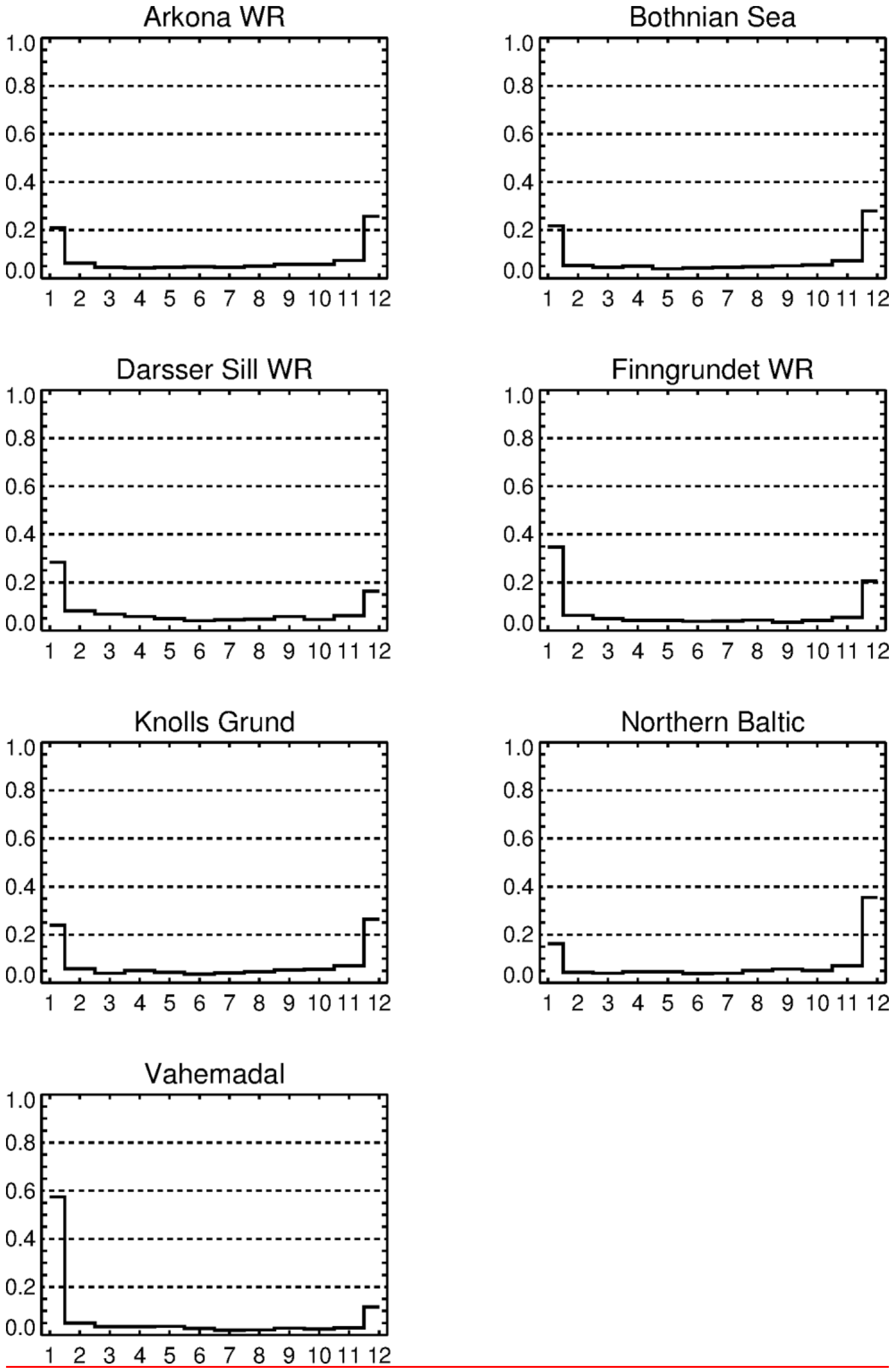
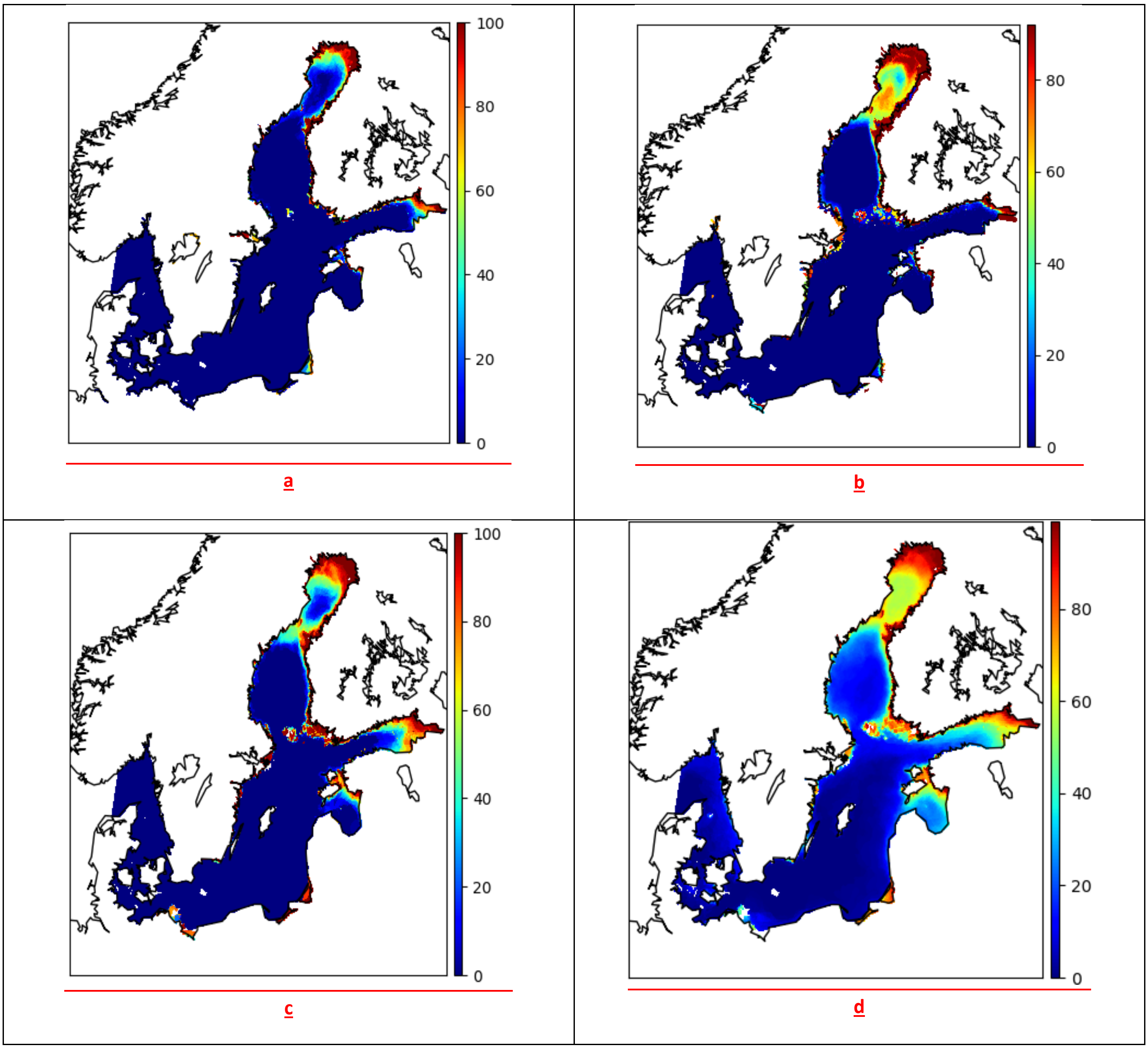


Figure S 11 As Figure S 10 but for forecast range 48 hrs.



a

b

c

d

Figure S 12 The average ice cover for February: a) 2015, b) 2016, c) 2017 and d) average 2010-2018.