Response to anonymous reviewer #1

Dear reviewer #1,

Thank you for your review and your comments. Additional supplemental material was prepared and uploaded regarding the calibration/validation procedure of the WWIII model and ensemble hindcasts of the storms Rafael and Toini. Please find in the following answers to your comments.

1 Major comments:

#1) The introduction is well written from the point of view of ensemble modelling, but it totally lacks material on Baltic Sea waves and the relevant research. Please see my list of references in the end as a starting point. Also the discussion of the results needs to be tied better to what we as a community know about Baltic Sea wave conditions.

Thank your for the list of publications concerning wave conditions in the Baltic Sea. We will include additional information and citations to existing studies in the introduction of the article. Following the article of Björkqvist et al. (2017) we added ensemble hindcasts of two additional storm events in the supplemental material and test another discretization of the energy spectrum, following Soomere (2005). This will be discussed in the article. We also add a short paragraph on Baltic Sea wave climate to the introduction.

#2) While I can get on board with using only one storm in this study, I think it is very unfortunate that the authors have chosen the 2002 storm when no data from the NBP wave buoy is available. For example the 2004 Rafael storm would have wave buoy data available for validation. It might be unreasonable to redo the model runs (I will leave that to the authors), but at least the authors should discuss how realistic the highest values (Hs>11 m) are by comparing to what we know about the Baltic Sea wave climate (again, see the list of references at the end).

Storm Rafael and Toini were additionally hindcasted with the newest setup (please see the supplemental material). WWIII was calibrated on basis of the UERRA/Harmonie-v1 wind data and gives a satisfactory perfomance (please compare the supplemental material). It is also shown that the wave heights for the two additional storms (Rafael and Toini) with both WRF-ARW and UERRA/Harmonie-v1 show realistic wave heights. For this reason, we assume that the significant wave height for the 2002 event with the unperturbed WRF-ARW wind forcing is also realistic. The perturbations of the WRF-ARW model physics were not tuned. To be able to do this, several extreme events would have to be hindcasted. One reason for the extreme wave heights in some ensemble members could therefore be an overdispersion of the wind fields from the WRF-ARW ensemble. In our WRF-ARW setup, the roughness length over the sea is assumed to be constant. Under severe storm conditions the sea surface roughness should increase with an effect on the wind field resulting in a limitation of the wave growth. A coupled WRF-WWIII setup would take this into account. By comparing the EPSgrams from the ECMWF (see for example ECMWF presentation [1] slide 20), one can see that the range of uncertainty can be very large. Based on a limited number of observations of extreme wave heights, it is therefore hard to judge which significant wave height is still realistic. We will discuss this in the article.

2 Specific comments:

#1) The wave model is "WAVEWATCH III", not "Wavewatch III"
This will be changed.

[1] https://confluence.ecmwf.int/download/attachments/55116817/OCEAN_WAVE_FORECASTING_AT_ECMWF_version_201602.pdf?api=v2

#2) page 1 line 25: Perhaps have a paragraph break at "In principle"?
A paragraph break is added there.

#3) page 3 line 23 "The ERA5 dataset was used in this study to drive the atmospheric model WRF, a coarse Wavewatch III wave model to provide lateral boundary conditions for a Wavewatch III wave model with higher resolution and for comparison with the model results." This is a bit unclear and should perhaps be rewritten.
Changed to: "ERA5 is used for the initial and lateral boundary conditions for the atmospheric hindcasts with the WRF model. Lateral boundary conditions for the Baltic Sea WAVEWATCH III setup originate from a setup for the North Sea. This coarser model is driven by ERA5 winds. ERA5 reanalysis and EDA data are used for comparison of the hindcasts produced with WRF and WAVEWATCH III."

#4) page 5 line 3 "UERRA/Harmonie-v1 was used for calibration and validation of the setup against one month of data from buoys available from the Copernicus Marine environment monitoring service 12 (CMEMS) with the previous Wavewatch III v5.16 version."
Information to theses questions is added in detail in form of supplemental material. At this part of the article, we will refer to this supplemental material. We see this article more as a demonstration of a principle idea for an ensemble hindcast procedure. For this reason, we think that it is sufficient if these details are presented as a supplement.

#5) Please add some kind of Table of the different type of ensembles. As written, it isa bit hard to follow.
Will be added.

#6) Fig 2: "results shown at 19.39°E, 56.17°N". Show this point in Fig. 1.
Will be added.

#7) Fig 3.There are a lot of subplot. Would it be sufficient to just use max difference to the mean,or to reduce the number of panels in some other way?
Figure 3 includes only the ensemble mean, minimum and maximum of the different ensemble generation approaches. Only the difference to the mean would neglect the fact that the spread cannot be assumed to be symmetric around the mean. Figure 4 includes also a lot of subpanels. This presentation called postage stamps, is often used to present ensemble forecasts. For this reason, we prefer to keep in this way.

#8) page 10 line 24: A shortcoming of this procedureA bit unclear what is meant by "this"
Changed to: "A shortcomig of the presented procedure for the wave hindcasts ..."

#9) page 11 lines 1-2: Baltic Sea not really swell dominated, so this shouldn't be an issue in your results, and the discussion seems a bit off key, especially in the middle of the paper concentrating on the Baltic Sea. It is up to the authors if they want to keep it. Just thought I would point out how it looks from a Baltic Sea perspective.
ERA5 is a global reanalysis. This is why the presented procedure for ensemble hindcasting can be applied for any region in the world. For this reason, we mentioned this point.

#10) page 13 line 2.Perhaps start a new paragraph with "Figure 8 shows..."?
We will start there a new paragraph as suggested.

#11) page 13 line 19 "The time step of a high resolution ocean or wave model is normally below one hour."This is slightly misleading, since one hour is a typical time resolution for the output of a wave model. The time step of a wave model can be counted in seconds (typical for explicit numerical schemes) or minutes (typical for implicit numerical schemes). The wave model therefore need updated wind information e.g. every 30 seconds. This is done by interpolation from the wind forcing that is provided e.g. every hour or every third hour.
We will adapt this part to: "The numerical time step of a wave model can be counted in seconds (typical for explicit numerical

schemes) or minutes (typical for implicit numerical schemes). The wave model therefore need updated wind information e.g. every 30 seconds. This is done by interpolation from the wind forcing that is provided e.g. every hour or every third hour."

#12) page 14 line 9-10: "Systematic differences cannot be found based on the small sample, but it indicates that the choice of the 15 minutes resolution is a reasonable compromise between a good representation of the extreme values and file size."I think one could argue that a 60 minute resolution is reasonable, since a difference of 2 cm is under 1%. This is small compared to the sampling variability (roughly 5-10%) that is present in measured significant wave height data that we routinely use to validate the models. Still, 15 minutes is clearly also a reasonable choice, so I'm not arguing with that part of your conclusion.
We agree that 60 minutes is reasonable. We only wanted to demonstrate that there might be an impact if using a higher temporal resolution. Of course, in the demonstrated case it is very small.

#13) page 14 line 15-16: "For this reason, a difference in the spatial pattern can beassumed. "Do you mean that a difference can be expected?
We change this to "expected".

#14) last paragraph on page 14: It think it is worth noting that the operational products typically used to force Baltic Sea wave models are already close to the higher resolution (0.063 deg). While this sensitivity test is very welcome, it could easily be read as if the wave modelling communityis currently using insufficient wind forcings is no context is provided. It might also beworth noting, that separate high-resolution wave model implementations might benefit more from higher resolutions in the wind forcing than what is seen in a 1 nmi BalticSea wide wave model. This kind of sensitivity tests for coastal wave models have been done in the Baltic Sea (see e.g. Tuomi et al., 2014).
This study has been done from the perspective of a research institute rather than an operational forecast centre. We are aware and mentioned it also in the manuscript that an operational product should be of higher quality then what we are able to do with this setup. As a research institution, we often do not have access or cannot rely on operational datasets only, since we are interested in hindcasting events over a long period as determined by the research question. We would be limited by applying operational products regarding to the available periods, but also in terms of homogeneity of the dataset, which is required for investigations of long-term changes. With ERA5 as a global reanalysis and the atmospheric and wave models available from github, we demonstrate an approach, which everybody could repeat for any region in the world. When ECMWF extends ERA5 back to 1950, nearly 70 years of data are available for the production of event based hindcasts in a homogeneous way. One very relevant question is then which resolution is neccessary and how large should be the ensemble for the hindcasts. Should we produce more members or do we get more benefit from a higher resolution ? We tried to discuss these issues in the article. The ensemble runs were also done here in a coarser resolution than 0.063 deg, because it would have delayed the study because of computational limits. We will include this point about the impact of higher wind field resolution on higher resolved wave models and will make it clear that the point of a refined horizontal resolution applies to hindcasts rather than operational applications.

#15) page 16 line 1-2: "As the first twelve hours are not used, because of the model spin-up, this is not really a shortcoming." This will not be true for operational wave forecasts that get their starting conditions fromthe previous run. Will it be a shortcoming then?
We use a reanalysis from a different model and coarser resolution as the WRF model. In an operational setup, one would probably use data assimilation which combines the background from a previous model run based on the same model with the same parametrisations with actual observations. This should reduce a spin-up significantly. There are other techniques to reduce the spin-up, also mentioned in the article, like Digital Filter Initialization for example. The spread develops also over the forecast horizon, why there might be a lack of spread during the first hours. This can be improved by applying an ensemble data assimilation technique.

#16) page 16 lines 11-13 "To achieve a comparable robust estimate of the uncertainty, the ensemble size for the here presented approach must be larger than the one of operational local area model ensembles. "Just to be clear, is the "here presented approach" choosing the members at random? In other words, is your conclusion that choosing random members requires more

members in the ensemble than if they are "screened" in advance using a coarse model, or are you trying to make some additional point?
With the presented approach, the ensemble size must be larger than in case of pre-selecting already a representative subsample of ensemble members, because the ensemble members are generated in a random way in terms of the stochastic perturbations. We will be more specific: "The here presented approach without pre-selection of ensemble member ..."

#17) page 16 line 16-17: "For a strong event, the difference between a 5 and 60 minutes temporally resolved wind forcing is only on the order of 2 cm. "I think it is a bit questionable to give an absolute difference without knowing the significant wave height. This doesn't really provide that much useful information.
The significant wave height of about 6.3m will be mentioned here.

#18) In e.g. Figure 2: are you using the wave product of ERA5, or are you using WAVEWATCH III forced with ERA5 winds?
We tested also the ERA5 wind as forcing data and found a relatively good model performance with an underestimation of the extreme wave heights in WWIII. For comparison, we showed the significant wave height from the ERA5 ECWAM with about 0.36° resolution (Fig. 6 and 7) and the ERA5 ECWAM uncertainty measure with about 1° resolution (Fig. 6). We will make this clearer.

#19) If you are only simulating the wave field in the Baltic Sea, then there is not really aneed to nest it outside of the Danish straits, since no significant amount of wave energywill penetrate. It's not wrong, just pointing out that it is not really necessary.
Our later application of the ensemble data are transport simulations with an ocean model for which we use the ensemble wave and atmospheric data as input fields. As we want to have also realistic wave parameters north of the Danish Straits, we used the presented nesting procedure.

#20) The figures are sometimes very hard to read. Please prepare them according tothe guidelines of the journal (fonts sizes, labeling of subpanels etc.
We will adapt the figures.

Response to anonymous reviewer #2

Dear reviewer #2,

Thank you for your review and your comments. Additional supplemental material was prepared and uploaded regarding the calibration/validation procedure of the WWIII model and ensemble hindcasts of the storms Rafael and Toini. Please find in the following answers to your comments.

1 Major comments:

This manuscript provides an interesting insight into possibilities of the construction of a large ensemble of hindcasts of wave properties in the Baltic Sea region. On the one hand, this approach is thought-provoking in itself as the pool of similar studies is very limited in this area. On the other hand, it is not clear beforehand how large is the potential of this approach to improve the hindcast as most of the discrepancies of the wave field reconstructions seem to stem from uncertainties of the driving wind fields. In particular, even small variations in the trajectories of low pressure systems may lead to large changes in the wave properties in the study area. It is thus important to understand how the possible uncertainties in wave reconstruction can be "distributed" between the variations in the driving fields and the specific ways of the description of wave physics. The topic thus clearly fits the scope of Ocean Science.

It is a pity that the approach is applied to an event in February 2002 for which essentially no ground truth about wave properties is available in the area of high waves. While thewave buoy of the Finnish Meteorological Institute was removed because of possible ice impact, the bottom-placed device at Almagrundet (Broman et al., 2006) did not provide any data in February 2002. However, as it is said both in Abstract and Conclusions that the event "provoked a severe storm surge in February 2002" it is necessary include at least some numbers and locations to substantiate this information. For example, nothing specific happened in Latvian waters.

Concerning the first remark about the applicability of this approach, it has to be mentioned that there is especially an interest from the insurance sector to produce large samples of historical events to get a more robust estimate of, for example, the 200 year return level as defined by the Solvency II directive. Often statistical methods are applied to enlarge the samples producible from datasets like reanalysis. Osinski et al. (2016) used the archive of EPS forecasts from the ECMWF to produce an enlarged ensemble of historical events. The problem with operational forecasts is the inhomogeneity and limited period. With our approach, ensemble hindcasts back to 1979 (eventually 1950 if ECMWF extends ERA5) can be created in a homogenous way. Our later application is a simulation of particle transport with an ocean model and a study of the impact of the metocean uncertainty on the transport pattern and amount of material.

Regarding your second remark about the missing observations, two storm events (Rafael and Toini) were hindcasted additionally. Information about the calibration procedure, validation of the model and the presentation of the two storm events is added in form of supplemental material. The results of the 2002 storm event were compared in the article to ERA5 wind and wave data. Based on a single event, it is not possible to judge if the ensemble spread is reasonable. For this reason, we compared it with the uncertainty measure provided with the ERA5 reanalysis to get a rough idea about it.

The method for the construction of the ensemble is rational and interesting. It is reasonable from the viewpoint of wind fields but seems to run into problems in terms ofwave properties. It is of course worth of trying to construct as large ensemble as possible in order to examine the spread. However, it is not a good sign that some membersof the ensemble lead to unrealistic wave heights. Both Fig. 2 and Fig. 3 indicate thatmaximum wind speeds in the northern Baltic proper are mostly in the range of 20–22m/s and only for a few members reach the level of 25 m/s. Such winds speeds onlycover a small part of the northern Baltic Proper. Even though the wind direction wasfavorable for the generation of high waves in this area, it is unlikely that significant waveheights substantially exceeded 7 m in this storm. Wave heights exceeding 8 m are veryinfrequent in this region. Even in the extreme storm Gudrun/Erwin (January 2005, 10-min wind speed >28 m/s in large sea areas) wave heights most likely did not exceed 10 m anywhere in the Baltic Sea (Soomere et al., 2008).

Therefore, I guess that wave heights between 11 and 12 m in Fig. 6 are completelyunrealistic for the February 2002 storm. It

seems that the entire ensemble severely(by almost 2 m on average) overestimates wave heights in the northern Baltic proper. Thus, I recommend to extensively comment this feature and to include a short insightinto measured or modelled wave heights in this area for storms of comparable prop-erties. Ideally, I would recommend to include a paragraph or two about extreme wave properties in the study area, following either (Tuomi et al., 2011) or (Björkqvist et al.,2018).

Calibration and validation of WWIII driven by the UERRA/Harmonie-v1 forcing dataset against observations and the hindcast of the additional two storm events showed that the waves predicted with the Baltic Sea setup with UERRA/Harmonie-v1 and the unperturbed WRF-ARW hindcasts show reasonable wave heights. Our WWIII setup was calibrated with UERRA/Harmonie-v1 forcing, for this reason it was checked whether this calibration gives also reasonable results when driven with WRF-ARW. The wind in WRF-ARW is slightly stronger over land and near the coast, as it can be seen in Fig. 4. We adapted the roughness length over land according to the Corine land cover data set, but this gives only a small effect for the waves in the western part of the Baltic close to the land masses. The roughness length over the sea surface is assumed to be constant in the applied WRF-ARW setup. Under severe storm conditions, the roughness of the sea surface should increase in reality, resulting in a reduction of the wind speed due to higher momentum transfer. Reduced wind speeds limit the growth of the wave height. With a coupled WRF-WWIII setup, this effect could be taken into account, in our setup it is neglected. Perhaps this is one reason for the extreme wave heights in some representations. Based on one extreme event, it is also not possible to tune the perturbations of the model physics. This is why the WRF-ARW ensemble is potentially overdispersive, which we also mentioned in the manuscript. As can be seen in Figure 4, the wind speed over the Baltic proper in the extremest representations is above 28m/s. The time series shown in Figure 2 is at a different location. We will make this clearer in the manuscript and will discuss it more in detail that the extreme representations are potentially unrealistic. As the two additional storms were hindcasted with a 7km newer WRF-ARW version, which we will apply for our later application, a recalculation of the 2002 storm shows a maximum hs of 9.5 m from an 11-member ensemble. The spread is larger than for the other two storm events which shows that the event is much more sensitive to perturbations.

In particular, I recommend extending the message on page 5, line 5–6 towards a sound explanation that the model is essentially uncalibrated for the Baltic Sea conditions. This is mentioned in the last sentence before conclusions on page 14. The point of this sentence should be made very clear from Abstract to Conclusions. I stress that such a bias in the evaluated wave heights does not undermine the validity of most of the results but it should be made clear to the reader that single values of wave height (and even the ensemble average) do not necessarily match the wave properties in this storm.

We will refer to the supplemental material and make it clearer that the WWIII setup was calibrated for the application with UERRA/Harmonie-v1, but a test with WRF-ARW wind also shows a reasonable performance. Concerning the extreme representation in some ensemble members, an additional discussion will be added about the uncalibrated ensemble spread in the WRF-ARW ensemble and about the fact that the effect of the roughness of the sea surface is not taken into account in the applied WRF-ARW setup.

For the listed reasons I recommend moderate to major modifications to the manuscript. It is essential that the reader is informed (i) about some basic features of wave climate and extreme waves in the Baltic Sea and also (ii) that the simulations probably strongly overestimate wave heights and (iii) are performed specifically to study the spreading properties of ensembles, with no exact relevance to the actual wave heights during the simulation interval. An absolute must is to inquire the modelled data from a properly calibrated run (e.g., from the authors of Björkqvist et al., 2018) for the underlying location of Fig. 6 to give a minimum flavor of the possible bias.

As proposed by both reviewers, additional information about the wave climate in the Baltic Sea including citations to existing studies will be included into the introduction together with the mentioned points from the previous remarks about the potential overestimation of the spread in the atmospheric ensemble data resulting in potentially unrealistic wave heights in some members. We believe that a more detailed comparison to observations makes an inter-model comparison no longer a requirement.

The text is written in fairly good English but reveals slight German accent in the form of very long sentences at places and missing of some articles in the text. It is mostly clear but still needs extensive polishing, especially closer to the end of the manuscript. As I am not native speaker, I only include a list of clear typos below.

We will revise the text.

2 Minor comments:

The paragraphs are at places very long. For example, the first paragraph of Introduction extends over 28 lines. It is recommended to split long paragraphs into shorter ones.
The paragraphs will be splitted into shorter ones.

The style of calendar days ("21. February 2002" on page 6, line 4 and "22nd to 24th of February" on the next line) should be unified.
The style of the calendar days will be unified.

The first two sentences of Abstract seem unnecessary
The second sentence explains issues in wave modelling and is required for the third sentence which claims that we address these by the presented method. The first sentence shall put the second one into context. We believe this sort of introduction is required to grasp the intention of the manuscript.

Page 1, line 17: probably should be "and is described".
Will be changed to "and is described".

Line 23 and some other locations: some journals require comma after "e.g."
Will be revised according to the requirement of the journal.

Page 2, lines 32–34: the sentence does not make sense; possibly because of too strong German accent.
Will be replaced by "' At the moment, the ensemble datasets in this project are limited in their temporal coverage or spatial resolution. It can be advantageous to be able to produce hindcasts of events whose spatiotemporal resolution is adapted to the requirements defined by a research objective."'

Page 3, line 13: C3S has already been explained on page 2, line 22.
Only the abbreviation will be used here.

Line 20: probably full stops are not necessary in "21. February 2002" and similar expressions.
Will be revised.

Line 23 it is better to say that 0.36deg and 1deg denote the resolution of the relevant grid. Please do so also in several locations below where the size in degrees is given without any explanation.
Will be adapted.

Page 4, line 10: please specify the meaning of "writing 15 minutes output".
Will be changed to: "and the model output interval is 15 minutes."

Line 12: please explain what is meant under "the temporal impact" (probably the dependence of the solution on the time step).
Will be changed to: "the dependence of the solution of the wave model on the temporal resolution of the wind data."

Line 17: please specify the meaning of "Eta layers".
Will be explained as a specific vertical coordinate system used for atmospheric models.

Line 18–19: consider replacing the jargon-like expression "until fine scales develop" by a more explanative one. Please do so also in several occasions below to avoid clash in the meaning of, e.g., "finer scales are not represented" on page 6, line 9.
We will replace "'scales"' by "'structures"' to avoid jargon.

Page 5, line 1: to avoid misinterpretation, I suggest to mention that nesting of the wave model to the Baltic Sea is not really necessary for the hindcast of wave properties in the central and northern regions of this water body because very little wave energy penetrates through the Danish straits.
The wave model output will be used as input for a model of the entire Baltic Sea which also covers a part north of the Danish straits. In this region, we also want to have reasonable wave parameters. This will be made clear in the manuscript.

The reasoning on lines 2–6 is only partially relevant for the conditions of the Baltic Sea.
It explains the procedure used for setting up the wave model. ERA5 is a global reanalysis and the procedure could be applied also to other regions in the world.

Line 7: while most of the model setup is obviously fine for the Baltic Sea, please com-ment on the adequacy of the use of the chosen frequency range for this water body. Wave modellers usually substantially extend the frequency space here. The team of the Finnish Meteorological Institute normally uses 35 frequencies (Laura Tuomi et al.,many papers) and some research in subbasins of the Baltic Sea even 42 frequencies (0.0418–2.08 Hz, Soomere, 2005). It is probably not necessary to cover such an extended range. However, insufficient coverage of short waves may lead to too slow wave growth under rapidly increasing wind conditions.
The discretization with 42 frequencies (0.0418–2.08 Hz, Soomere, 2005) together with a finer resolution of the directions (36 every 10deg) was tested. In the supplemental material, the outcome is visible. It brings additional 10cm in the significant wave height for the Rafael storm, which is underestimated by about 90cm with the UERRA/Harmonie-v1 wind. The shortcoming of this finer discretization is a prolongation of the calculation time, which was 4 times of the one with the ERA5 equivalent discretization. For computational reasons, we used the ERA5 discretization.

Lines 14–16: the message of the entire sentence is technically clear but seems misplaced or even irrelevant.
Will be revised.

Page 7, line 7: "these".
Adapted.

Page 10, lines 19–20, the sentence "Compared to ERA5, the overall spatial pattern is comparable" does not make sense to me.
Will be replaced by "'The overall spatial pattern of the significant wave height is comparable between ERA5 and the WRF ensemble members."'

Page 11, lines 2–5: the reasoning is almost irrelevant for the Baltic Sea conditions andshould be left out. Instead, it should be emphasized that strong swells are infrequentin the Baltic Sea (see, e.g., Broman et al., 2006; Soomere et al., 2012) and thus deviations in the hindcast or forecast driven by the accuracy of the representation of swells are usually not very large in this water body.
We see the manuscript as a demonstration for the procedure to produce ensemble hindcasts. ERA5 is global and the procedure is applicable in general worldwide. This is why we also have to mention potential shortcomings if applying the procedure to other regions. We will add a subsentence "', which should, however, be more relevant for different regions of interest where swell plays a larger role."'

Page 12, line 14: something is wrong with "500 choose N possibilities exist".
This is an expression from stochastics, we will replace it by the mathematical notation $\binom{500}{N}$ to avoid confusion.

Page 13, line 10–12: the sentence is unclear.
Will be revised.

Line 13: "developed"; also, the entire sentence remains partially unclear starting from "why".
Will be revised.

Lines 16–17: the concluding sentence of the subsection should be made clearer.
Will be changed to "'Depending on the application, the ensemble size needs to be selected by a compromise between the robustness of the uncertainty estimate and the computational cost."'

Line 18: use "on" instead of "onto".
Will be changed.

Page 14, line 7: please specify what is meant under "The higher temporal resolutions do not differ so much." Also, the subsequent sentences contain too much jargon.
Will be revised.

Line 14: "orography of the coastlines" sounds weird as the height of the coastline is just zero; also: use "Baltic Sea".
Will be revised.

Line 15: spatial pattern of what?
of the significant wave height.

Line 5 or another appropriate place: please stress that an uncalibrated (for the Baltic Sea conditions) wave model was used but still the results about the spread are valid.
We will refer to the supplemental material and the issue with the spread will be discussed there.

Page 16, line 1: remove "by this fact".
Removed.

Lines 1–2: the message of the sentence "As the first twelve hours are not used, be-cause of the model spin-up, this is not really a shortcoming." remains unclear.
Will be revised.

Line 14: correct "atmopsheric".
Corrected.

Line 20: correct "possbile".
Corrected.

Ensemble hindcasting of wind and wave conditions with WRF and ~~Wavewatch~~ WAVEWATCH III® driven by ERA5

Robert Daniel Osinski[1] and Hagen Radtke[1]

[1]Leibniz Institute for Baltic Sea Research Warnemünde, Seestrasse 15, 18119 Rostock, Germany

Correspondence: Robert Daniel Osinski (robert.osinski@io-warnemuende.de)

Abstract. When hindcasting wave fields of storm events with state-of-the-art wave models, the quality of the results strongly depends on the meteorological forcing dataset. The wave model will inherit the uncertainty of the atmospheric data, and additional discretisation errors will be introduced due to a limited spatial and temporal resolution of the forcing data. In this study, we demonstrate that applying an atmospheric downscaling with the atmospheric mesoscale model WRF can address all these three issues. Not only does it add regional detail to the wind field and can increase the temporal resolution of the wind fields, which gives a more detailed representation of transient events such as storms. It can also be used to generate ensembles with perturbed atmospheric conditions which allow for a flow dependent and spatiotemporally variable uncertainty estimation. We test different strategies to generate an ensemble hindcast of a storm event in February 2002 in the Baltic Sea, which provoked a severe storm surge. The WRF model used for this purpose is driven by the ECMWF ERA5 reanalysis, and wind fields are passed to the third-generation wave model ~~Wavewatch~~ WAVEWATCH III®. A combination of initial conditions from the ERA5 ensemble of data assimilations and stochastic pertubations during runtime is identified as the most promising strategy. The final aim of the ensemble approach is to quantify the hindcast error, but this approach can also be used to generate alternative representations of historical extreme events to sample the recent climate and to increase the sample size for statistical studies, such as for civil engineering applications for coastal protection studies.

1 Introduction

The Lorenz attractor (Lorenz, 1963) is often used as an example to motivate ensemble forecasts. It explains a chaotic system behaviour, which is very sensitive to slight differences in the initial conditions and ~~it~~ is described by a system of differential equations. In operational weather prediction, ensemble forecasts are a common tool to quantify the forecast uncertainty by producing a set of alternative realisations. Initial conditions are estimated by data assimilation combining observations with a background field, which is normally a previous model run. The sparse spatiotemporal observational coverage leads to uncertainties in the initial conditions, which are growing over the integration time. A second type of uncertainty comes from the model parametrisations. These are used to take processes into account, which cannot be resolved by the dynamical core of the model, e.g. subgrid-scale processes like turbulence or convection, or processes which can be described physically but are computationally too expensive to explicitly take them into account (e.g. utilisation of a 1-moment instead of a 2-moment micro physics scheme).

In principle, three methods exist to generate an ensemble forecast~~. Some of them are tested~~ , the latter two of which are these are tested in this communication to estimate the uncertainty of a hindcast. ~~One~~ The first possibility is the combination of forecasts from different models (e.g. Hagedorn et al., 2005) or using the same model with different types of model physics (e.g. Ricchi et al., 2019). This multi-model/physics approach has the disadvantage that the ensemble size is limited to the number of available models/physics packages. Also, the forecast skill over a specific region and a specific variable might differ between the different models, what has to be taken into account in the interpretation. A second approach is the combination of forecast runs from the same model for the same time instance, but started at different initialisation times, called lagged-average forecast (LAF) ensemble (Hoffman and Kalnay, 1983). A limitation here is also the number of forecasts covering the same time instance and the fact that a newer forecast can be expected to have in average a better forecast skill than a forecast at long lead times. The third method is the utilisation of a single model and applying pertubations to the initial conditions and/or to the model physics.

Such an approach is used operationally at the European Centre for Medium-Range Weather Forecasts (ECMWF) since 1992. Initial conditions are perturbed by singular vectors (Buizza, 1998) or by a combination of Ensemble data assimilation (Buizza et al., 2008) with singular vectors, or by breeding vectors (Toth and Kalnay, 1997) like in case of the National Centers for Environmental Prediction (NCEP). Stochastic perturbations like Stochastically Perturbed Parametrization Tendencies (SPPT) (Buizza et al., 1999), Stochastically Perturbed Parametrizations (SPP) (Ollinaho et al., 2017) and Stochastic Kinetic Energy Backscatter (SKEB) (Shutts, 2005) are used to perturb the model physics (Leutbecher et al., 2016, 2017). SPPT perturbs the model parametrizations by applying a multiplicative noise and SKEB simulates the upscale transfer of kinetic energy from smaller to larger scales. Besides the application of SPPT in the global ECMWF medium-range ensemble system, stochastic perturbations are also used in local area models (e.g. Bouttier et al., 2012) and in ocean models like in NEMO (e.g. Brankart et al., 2015).

In a well constructed ensemble, the ensemble spread reflects the average forecast error. Stochastic perturbations need some time until a reasonable spread develops. Ensemble data assimilation (EDA) gives different estimations of the initial state representing its uncertainty. A forecast started from the different members develops the desired ensemble spread faster.

ERA5 (Copernicus Climate Change Service (C3S), 2017) is the newest global reanalysis from ECMWF. The resolution is relatively high with about 31 km resolution for the atmospheric variables, but depending on the application, it can be still too coarse. From ERA5, in contrast to previous reanalyses, an uncertainty measure based on an ensemble of data assimilation is available.

The Weather Research and Forecasting (WRF) (Skamarock et al., 2019) model is widely used in research as well as in operational weather forecasting and includes implementations of the mentioned stochastic perturbation schemes. The motivation of driving WRF with this new dataset is to be able to produce hindcasts of atmospheric conditions in different spatiotemporal resolutions including a measure of uncertainty based on ensemble techniques. This allows for example to study the effect of the model resolution on effects like up- and downwelling in coastal regions.

Some regional reanalysis (ensemble) datasets are already freely available. Such regional reanalyses are produced, for example, in the framework of the project "Uncertainties in Ensembles of Regional ReAnalysis" (UERRA)[1]. ~~The~~ At the moment, the ensemble datasets in this project ~~cover at the moment only short periods or have a coarse resolution, why it~~ are limited in their temporal coverage or spatial resolution. It can be advantageous to be able to produce hindcasts of events ~~during a longer period, allowing also to adapt the spatiotemporal resolution to the needs~~whose spatiotemporal resolution is adapted to the requirements defined by a research objective. It has to be mentioned that the quality of a freely running hindcast can be expected to be inferior to such a re-/analysis product containing state-of-the art data assimilation techniques. Another database from which ensemble forecasts of local area models are available is from the Tigge-LAM archive[2] (Swinbank et al., 2016). The available forecast models cover also only short periods and they are operational, meaning that the datasets are not homogeneous, because the model version can change during time.

The Baltic Sea, which is a marginal sea in the north-east of Europe, is taken as an example for the application of the demonstrated procedure to produce ensemble hindcasts of wind and wave conditions by driving the WAVEWATCH III® wave model with wind data produced with the WRF ensemble model. Observed wave heights in this region do not exceed 8.2 m (Björkqvist et al., 2017) and waves are dominated by the wind sea (Broman et al., 2006; Soomere et al., 2012). More detailed information about the Baltic Sea wave climate for specific subregions is provided, for example, by Björkqvist et al. (2017), Soomere (2005), Soomere et al. (2008), Tuomi et al. (2011) and Tuomi et al. (2014). As ERA5 is a global reanalyis, the demonstrated procedure is also applicable in other regions.

The idea behind this study is to generate an ensemble hindcast on event basis in a comparable way to operational weather forecasts by driving WRF with ERA5 including the initial conditions from the ERA5 EDA with stochastic perturbations (SKEB and SPPT). Other ensemble generation techniques are tested for comparison. The atmospheric data from a hindcast or forecast are discrete in time and space. This limits the accuracy and affects the outcome if driving another model like an ocean or wave model. This uncertainty is investigated by driving ~~a~~ the wave model with different spatiotemporal resolutions.

2 Data and models

2.1 Data

2.1.1 ERA5

ERA5[3] ~~(Copernicus Climate Change Service (C3S), 2017)~~ (C3S, 2017) is the follow-up ECMWF reanalysis of ERA-Interim produced with the Integrated Forecasting System (IFS) cycle 41R2[4], operationally at ECMWF in March 2016. It is provided under the Copernicus licence[5] allowing also commercial applications. Hourly reanalysis in about 31 km (∼0.28°deg) horizon-

[1] http://www.uerra.eu/

[2] https://apps.ecmwf.int/datasets/data/tigge-lam/expver=prod/type=pf/

[3] https://confluence.ecmwf.int/display/CKB/ERA5+data+documentation

[4] https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation

[5] http://apps.ecmwf.int/datasets/licences/copernicus/

tal resolution and 137 vertical model levels are available from 1979 (eventually 1950) and the dataset is getting prolongated into the future with a delay of about three months. A state-of-the-art data assimilation technique is used (4D-Var). ~~Additionally~~ In addition to the reanalysis, on three hourly basis, ten members of an ensemble of data assimilation (EDA) are provided as an uncertainty measure with half of the resolution of the reanalysis. The reanalysis data of surface fields and the 137 model levels were extracted on hourly basis interpolated onto a slightly higher 0.25~~°~~ deg resolution grid for the period ~~21.~~ 21 February 2002 until ~~24.~~ 24 February 2002 as recommended by ECMWF. ERA5 data from the ensemble of data assimilation were also interpolated bilinearly onto the same 0.25~~°~~ deg regular longitude-latitude grid. ERA5 also includes fields from the ECWAM wave model (ECMWF, 2016) in 0.36~~° and in 1°~~ deg and in 1deg from the ensemble of data assimilation. ~~The~~ ERA5 ~~dataset was used in this study to drive the atmospheric model WRF , a coarse Wavewatch III® wave modelto provide lateral~~ is used for the initial and lateral boundary conditions for the atmospheric hindcasts with the WRF model. Lateral boundary conditions for ~~a Wavewatch III® wave model with higher resolution and for comparison with the model results~~the Baltic Sea WAVEWATCH III® setup originate from a setup for the North Sea. This coarser model is driven by ERA5 winds. ERA5 reanalysis and EDA wind and wave data are used for comparison of the hindcasts produced with WRF and WAVEWATCH III®.

2.1.2 UERRA/Harmonie-v1

The UERRA/Harmonie-v1 dataset (Ridal et al., 2017) contains analyses at 00, 06, 12 and 18 UTC as well as hourly forecasts for +1h until +6h and thereafter ~~three hourly~~ three-hourly until thirty hours. The Harmonie model is used for the production of this dataset in about 11 km horizontal resolution and 3D-Var data assimilation is used with conventional observations (synoptic stations, ships, drifting buoys, aircraft observations and radio soundings). Large scales from ERA40 and ERA-Interim are introduced into the data assimilation by large scale mixing. The available period extends back until 1961. To create an hourly dataset, the analysis fields were combined with the forecast lead times +1h to +5h, retrieved from ECMWF[6]. Wind data were interpolated bilinearly onto the regular wave model grid for the Baltic Sea described in the next section. This dataset was mainly used to produce a restart file for the wave model runs and for calibration/validation of the wave model.

2.2 Models

2.2.1 Atmospheric Weather Research and Forecasting model (WRF)

The Weather Research and Forecasting model WRF v4.0.3 model[7] in the Advanced Research WRF (ARW) ~~(Skamarock et al., 2019) version~~ version (Skamarock et al., 2019) is applied here. It is used in non-hydrostatic mode in 0.126~~° horizontal resolution writing~~ deg horizontal resolution and the model output interval is 15 minutes~~output~~. To investigate the ~~impact of different horizontal resolutions on the significant wave height,~~ dependence of the solution of the wave model on the spatial and on the temporal resolution of the wind data, runs in 0.252°deg and 0.063~~°~~ deg were produced as well as output at a temporal resolution of 5 minutes~~to investigate the temporal impact.~~ . In this way, a factor of about 4.5 between the highest WRF resolution and the

[6]https://apps.ecmwf.int/datasets/data/uerra
[7]https://github.com/wrf-model/WRF/releases

driving ERA5 fields is given, and the same factor between the ERA5 EDA fields and the WRF ensemble runs. The domain is slightly larger than the Baltic Sea for all runs. For the model configuration, the CONUS physics suite (Wang et al., 2019) is used. This is a combination of model physics adapted for the Continental United States of America. As it is well tested, this physics setup is taken as it is, and we assume that it should be reasonable for other regions in the mid-latitudes. The 89 vertical Eta layers used in this WRF setup, a specific vertical coordinate system in atmospheric models, are adapted to be comparable to layer 2 to 90 of the IFS[8] until 50~~hPa~~ hPa. As initial conditions come from a different and coarser model, it needs some time until fine ~~scales~~ structures develop. Methods for spin-up reduction like Digitial Filter Initialisation (Peckham et al., 2016) have not been tested. Instead, the WRF output is only used twelve hours after initialisation to drive the wave model. Neither data assimilation nor observation nudging is used. Hindcasts are produced in this study in a comparable way like a forecast, downscaled from a global forecast model. For this reason, the results of this study are valid for both hindcasts and such forecasts. The WRF Pre-Processing System (WPS) in version 4.0.3 is used to prepare the input data for the model together with the WPS V4 Geographical Static Data[9].

2.2.2 Wave model ~~Wavewatch~~ WAVEWATCH III®

~~Wavewatch~~ WAVEWATCH III v6.07® [10] (Tolman, 1991; The WAVEWATCH III® Development Group (WW3DG), 2019) is used in this study for the Baltic Sea. It is a state-of-the art third generation wave model, which is also used as an operational wave forecast model. A one-way nesting approach is applied, see ~~figure~~ Figure 1: A setup with 0.1~~°~~ deg resolution covering the North Sea and a small part of the eastern Atlantic ocean is used to produce boundary conditions for the Baltic Sea setup at the border with the North Sea. This is not really necessary for the central and northern regions of the Baltic Sea, as very little wave energy passes the Danish straits. To avoid showing unrealisic values in a part of the domain, the nesting procedure was nevertheless applied. The GEBCO_2014 Grid in version 20150318[11] is used as bathymetry. The Baltic Sea setup has a resolution of one nautical mile with 149.282 sea grid points and the bathymetry is based on the work of Seifert et al. (2001). UERRA/Harmonie-v1 was used for calibration and validation of the setup against one month of data from buoys available from the Copernicus Marine environment monitoring service[12] (CMEMS) with the previous ~~Wavewatch~~ WAVEWATCH III v5.16 version. A calibration and validation with the WRF forcing was not yet possible because of the short period that has been hindcasted until now. Nevertheless, the wave model shows a satisfactory performance with the WRF forcing. Detailed information about the calibration and validation procedure of the wave model can be found in the supplemental material. 24 directions starting at 7.5° with a 15° direction increment and 30 frequencies starting at 0.03453 Hz geometrically distributed with a step of 1.1 are used for the discretisation of the energy spectrum. This is comparable to the settings for the wave model in ERA5. Soomere (2005) proposes a finer resolution of the energy spectrum. This finer resolution was tested

[8] https://www.ecmwf.int/en/forecasts/documentation-and-support/137-model-levels
[9] http://www2.mmm.ucar.edu/wrf/src/wps_files/geog_high_res_mandatory.tar.gz
[10] https://github.com/NOAA-EMC/WW3
[11] http://www.gebco.net
[12] http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=INSITU_BAL_NRT_OBSERVATIONS_013_032

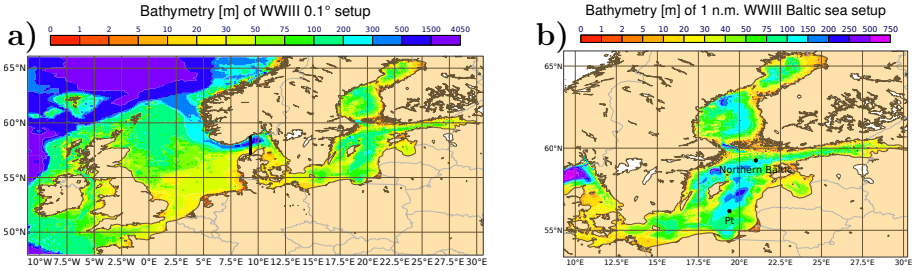

Figure 1. Bathymetries [m] and domains of a) 0.1° deg and b) 1 nautical mile ~~Wavewatch~~ WAVEWATCH III® setups; in black ~~points~~ in the left panel ~~show~~ a) grid cells for the nesting of the Baltic Sea model are shown. The black point "Pt" in the right panel b) shows the location of ~~timeseries for figure~~ the time series in Figure 2 and the black point "Northern Baltic" the location of the time series in Figure 6.

and the result is demonstrated in the supplemental material. A clear impact on the extreme wave heights is visible, but it prolongates significantly the computing time. For our specific application, the ERA5 discretization is a good compromise between computational effort and model performance. The physics packages are defined before compiling the model by a so-called switch file. The switch file Ifremer1, provided with the model code, is applied in this study. This includes wind input and dissipation after Ardhuin et al. (2010) and the SHOWEX bottom friction scheme (Ardhuin et al., 2003). A sediment map based ~~from~~ on the European Marine Observation and Data Network EMODnet[13] data ~~were~~ was used for applying non-homogeneous bottom friction. The model runs were produced between ~~2002-02-22 00UTC and 2002-02-24 00UTC. Wavewatch~~ 22 February 2002 00 UTC and 24 February 2002 00 UTC. WAVEWATCH III® was started from initial conditions from a previous run conducted ~~over~~ for 21 days driven with UERRA/Harmonie-v1. The sea ice area fraction is taken from ERA5. In the atmospheric model, the stochastic perturbations of the model physics contribute significantly to the ensemble spread. Wave models include different source terms (e.g. wave generation, dissipation, bottom-friction, and so on), which are partly simplified to make the model computationally more efficient or are described empirically (Farina, 2002; Yildirim and Karniadakis, 2015). Nevertheless, the wave model ensemble approach here is based solely on the ensemble of the atmospheric forcing data and includes no perturbations of the source terms.

3 Ensemble hindcasts

3.1 Wind fields

Six different approaches to generate an ensemble hindcast are presented in this section, see Table 1. The first approach is to generate an LAF ensemble. This is done by initialising the WRF model at different times on ~~21.~~ 21 February 2002 at every hour between 08 and 16 UTC, which results in 9 runs covering the period from ~~22nd to 24th of February .~~ 22 to 24 February 2002. The second approach is based on the domain shifting presented by Pardowitz et al. (2016). The ERA5 reanalysis is for this purpose shifted by one grid cell (0.25°deg) in each direction horizontally producing 8 perturbed ensemble members. For

[13] http://www.emodnet-geology.eu/

Table 1. Methods tested for the generation of an ensemble hindcast with WRF

Method	Procedure	No. of members
1	LAF approach, WRF initialised at different times (21 February 2002 between 08 and 16 UTC)	9
2	Domain shifting approach, ERA5 shifted horizontally by one grid cell	8
3	ERA5 EDA fields used for initial conditions	10
4	Stochastic perturbations (SKEB and SPPT) together with random perturbations of LBC's	10
5	As approach 4, but initialised from ERA5 EDA as in approach 3	10
6	As approach 5, but additional runs started at three hours earlier and later	30

the third approach, WRF is initialised from the ERA5 fields from the ensemble of data assimilation. These fields have a coarser resolution, but they are used in this study as the ERA5 reanalysis in 0.25°deg. This has the disadvantage that finer scales are not represented, but this is comparable to a downscaling from a global ensemble model, except that the reanalysis is used here as lateral boundary condition. As an alternative to keep the finer scales, it was tested to add perturbations to the initial fields, calculated by the difference between the ERA5 EDA members and the EDA ensemble mean, once with positive and negative sign to the ERA5 HRES reanalysis. We didn't find an improvement against the direct application of the ERA5 EDA fields. SKEB and SPPT are used for the fourth approach and the fifth combines approach three and four. For approach six, the same setup is used as in approach five, but as the ERA5 EDA fields are available every three hours, runs three hours earlier and later are additionally used as in an LAF approach. This leads to a thirty member ensemble.

In an ensemble system, it is important that the ensemble spread reflects the uncertainty. If the spread is too narrow, the system is underdispersive meaning that the forecast is overconfident and vice versa for an overdispersive/underconfident forecast. One tool for quantifying the quality of the ensemble spread is, for example, the Talagrand (rank) diagram (Hamill, 2001), and there are other quality measures like for example reliability, resolution, accuracy or sharpness, which are important for a good ensemble system (Murphy and Winkler, 1992). To be able to use the traditional ensemble verification methods (Jolliffe and Stephenson, 2003), long time series are needed, which could not be produced in this study. For this reason, an absolute statement which of the tested approaches performs best can not be given based on only one single hindcasted event. The different approaches are compared against the ERA5 reanalysis and the ERA5 members from the ensemble of data assimilation. As a larger variability can be expected in the higher-resolution model, it can be assumed that it increases also the uncertainty, what should be reflected by a larger spread than found in the much coarser data from the ERA5 ensemble of data assimilation.

A good agreement at a specific location between the ERA5 reanalysis and the WRF runs is visible during the first twenty hours in ~~figure~~ Figure 2. The wind speed maximum is higher than in ERA5. For comparison, the closest grid cell of the UERRA/Harmonie-v1 data is plotted and also shows higher values than ERA5. From ERA5, also the wind speed from the closest grid cell of the 0.25° deg grid is plotted, why some deviations from the initial conditions, which were prepared differently with the WRF preprocessing system, are visible. The resolution of the WRF runs is closer to the one from UERRA/Harmonie-

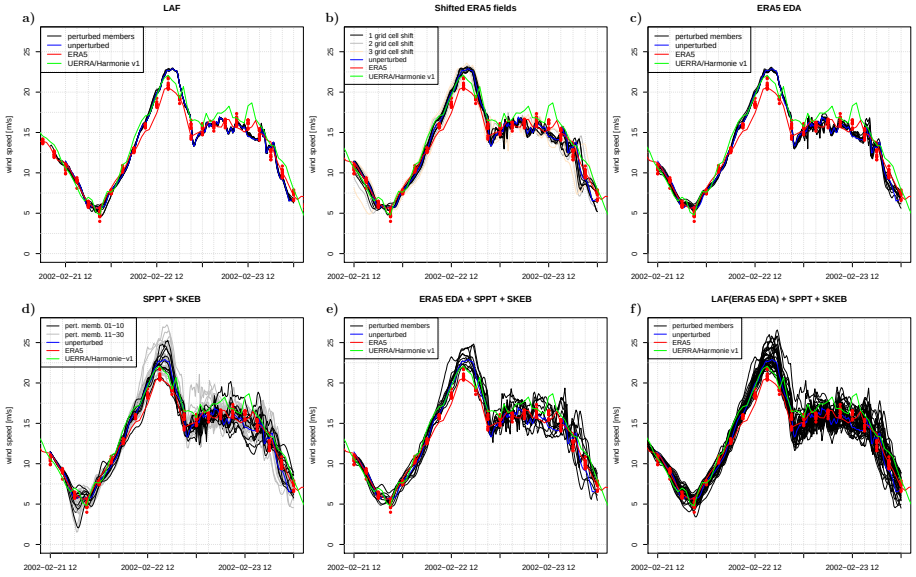

Figure 2. Time series demonstrating the simulation results of the different ensemble generation strategies at one specific location: a) lagged-average forecast (LAF) ensemble (Hoffman and Kalnay, 1983), b) domain shifting (Pardowitz et al., 2016), c) WRF runs started from ten ERA5 4D-EnVAR members with HRES LBC's, d) stochastic perturbations (SPPT and SKEB) (Buizza et al., 1999; Shutts, 2005), e) ERA5 4D-EnVAR as starting conditions plus stochastic perturbations, and d) LAF started from ERA5 4D-EnVAR at 09, 12 and 15 UTC plus stochastic perturbations; results shown at 19.39°E, 56.17°N

v1, and a stronger variability and also higher extremes can be expected due to the difference to the ERA5 resolution. WRF adds additional information from the finer scales and resolves the orography, coastlines and islands better.

All ensemble techniques lead to deviations from the unperturbed run. The LAF ensemble shows a very small spread. In fact, this is good, because it means that irrespective of the starting time of the WRF model being shifted by a few hours, the outcome is comparable. The first three approaches show a lower ensemble spread than the ensembles which include stochastic perturbations. Compared to the ERA5 EDA members, it gives an indication that these ensembles could be underdispersive. During the first hours of the ensemble with only stochastic perturbations (ensemble approach 4), all members are identical, as it needs some time until the perturbations introduce spread. Starting from the ERA5 ensemble of data assimilation (ensemble approaches 3, 5, and 6), spread is visible from initialisation on. Even with the coarser resolution of ~~theses~~ these fields, its application seems to be working, but additionally stochastic perturbations are necessary to produce a larger spread. The WRF ensemble started from ERA5 EDA fields at 09, 12 and 15 UTC also represents the uncertainty at ~~February 21st at 21UTC~~21 February 2002 at 21 UTC, where the lowest values in the ERA5 EDA members (Fig. 2) in the shown period can be found. With only stochastic perturbations, such low values are also visible, but a few hours too early. For the last simulation day, the spread of the combined ERA5 EDA and stochastic perturbations approach is very large, but it could not be tested if it is overdispersive.

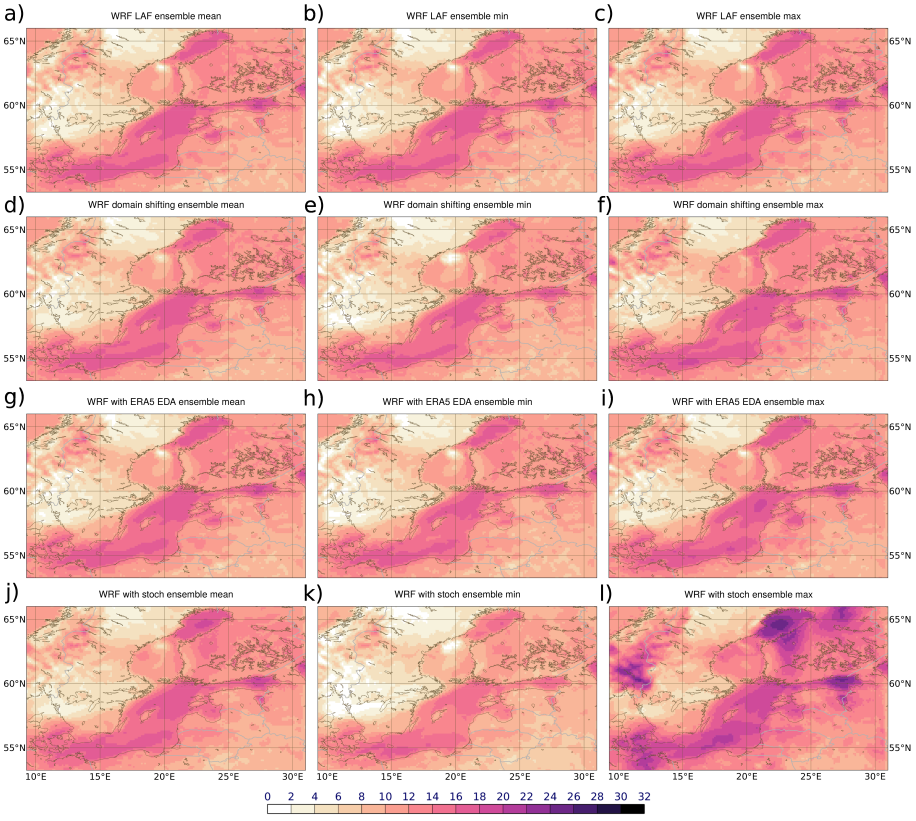

Figure 3. Ensemble mean a,d,g,j), minimum b,e,h,k) and maximum c,f,i,l) wind speed [m/s] of WRF ensemble based on LAF approach a-c), domain shifting d-f), initial conditions from the ERA5 EDA g-i) and on j-l) stochastic perturbations. All initialised at ~~2002-02-21 12UTC~~21 February 2002 12 UTC. Shown ~~2002-02-23 09UTC~~23 February 2002 09 UTC.

Spatially (Fig. 3), the spread in the WRF ensemble started from ERA5 EDA is very small. A much larger spatial variability ~~is presented~~ appears by applying stochastic perturbations. Especially strong wind is present in some members over the northern part of the Baltic Sea. The LAF approach also shows very little spread spatially over the entire domain. Domain shifting also did not produce as strong variability as applying stochastic perturbations. The combination of ERA5 EDA and stochastic perturbations produces members which show a strong variability in the central Baltic Sea (Fig. 4).

The LAF approach also shows very little spread spatially over the entire domain, and domain shifting also did not produce as strong variability as applying stochastic perturbations (Fig. 3).

A strong variability in the Northern as well as in the Central Baltic Sea is present by initialising WRF at 09, 12 and 15 UTC from ERA5 EDA fields with stochastic perturbations. Ten members are a small number to sample the uncertainty. Comparing a ten with a thirty member ensemble is not really a fair comparison. Ensemble approach number 6 shows in the ensemble maximum of all ensemble members high values in the central as well as in the northern part of the Baltic Sea. Figure 2 shows also the WRF ensemble with only stochastic perturbations and thirty members. The spread is in this case larger, but still inferior

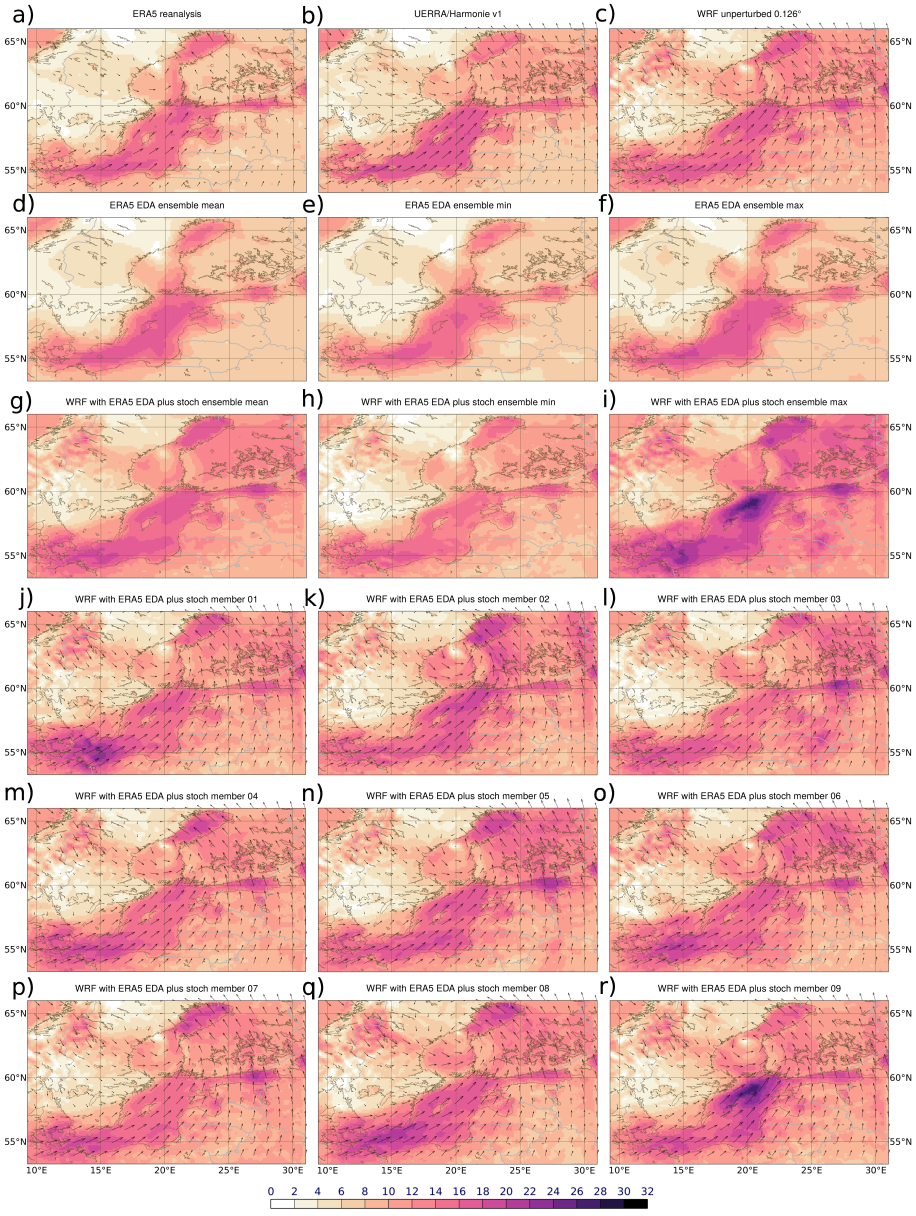

Figure 4. Postage Stamps: WRF ensemble approach number five generated by starting the ten members from ten ERA5 EDA members at ~~2002-02-21 12UTC~~ 21 February 2002 12 UTC plus stochastic perturbations SPPT and SKEB. Shown ~~2002-02-23 09UTC~~23 February 2002 09 UTC. a) The ERA5 reanalysis, ERA5 EDA ensemble d) mean, e) minimum and f) maximum, c) WRF unperturbed, WRF g) ensemble mean, h) minimum and i) maximum, b) UERRA/Harmonie-v1, and ~~the~~ nine j-r) perturbed WRF members are shown. Wind ~~Speed~~ speed [m/s] and direction as arrows.

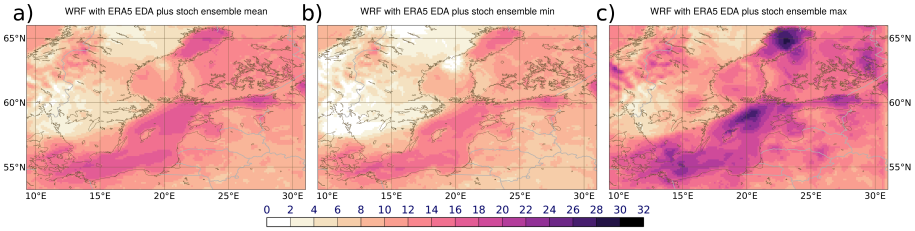

Figure 5. Ensemble mean a), minimum b) and maximum c) wind speed [m/s] from WRF ensemble approach number 6 generated by starting three times ten members from ERA5 EDA members at ~~2002-02-21~~ 21 February 2002 09/12/15 UTC plus stochastic perturbations SPPT and SKEB. Shown ~~2002-02-23 09UTC~~23 February 2002 09 UTC.

to the thirty member approach number 6 with ERA5 EDA as initial conditions and stochastic perturbations. Also the region in the central Baltic Sea gains spread by adding additional members, but contains lower spread than in approach number 6 shown in ~~figure~~ Figure 5. This demonstrates that ten members might be still not sufficient to sample the entire range of uncertainty, and that the combined application of model and initial perturbations is beneficial to create a larger spread.

3.2 Wave fields

The LAF and domain shifting approaches were not used to drive the wave model, because they show a relatively small spread. Figure 6 shows a time series at a station in the central Baltic Sea (see ~~figure~~ Figure 1). The comparison with the closest grid cell from ERA5 shows a good agreement in the temporal evolution of growth and a comparable trend in the decay of the significant wave height, but the maximum peak is about one metre lower in ERA5. ERA5 also shows the second peak only very weakly and some hours later during the middle of the second simulation day. ~~Wavewatch~~ WAVEWATCH III® in this study has a much higher resolution with 1 n.m. compared to the 0.36~~°~~ deg ECWAM model of the ERA5 reanalysis and the WRF wind forcing is spatially (0.126~~°~~ deg vs. 0.28~~°~~deg) and temporally (15' vs. 60') of higher resolution. This can explain locally much higher values and a stronger variability. Especially the maximum of the significant wave height varies strongly between the different ensemble realisations. This difference can be due to a different dynamical evolution of the storm or due to different tracks in the atmospheric model members (compare Osinski et al., 2016), leading to differences in the position. Already a slight change in the track of the storm can provoke large differences in the maximum if looking at a specific location in such a high resolution. With 0.36° resolution in ERA5, a slight change in the track can be assumed to not lead to such strong differences.

Figure 7 shows the different wave model members driven by WRF with ERA5 EDA initial conditions and stochastic perturbations. All members show high values in the central Baltic Sea. The time series shown in ~~figure~~ Figure 6 is in this region with the highest significant wave height on ~~22.~~ 22 February 2002 at 09 UTC. There is also a strong variability between the different ensemble members in this region. Wave heights in member 8 are especially higher than in the other members in the Gulf of Bothnia, but this member shows also much higher wave heights than the other members in the central Baltic Sea. In the western Baltic Sea the differences between the members are not that strong. ~~Compared to ERA5, the~~ The overall spatial pattern ~~is comparable~~of the significant wave height is comparable between ERA5 and WRF ensemble members. The wave

models (Fig. 6) driven by the WRF ensemble hindcast started from the ERA5 EDA initial conditions show a very small spread. A difference can be especially seen at the second peak. Much stronger differences are provoked by the WRF ensemble based on stochastic perturbations. Combining both ERA5 EDA fields as initial conditions and stochastical perturbations produces a comparable spread. The simulated significant wave heights of the most extreme members with about 11.2 m are clearly above the highest observations with about 8.2 m (Björkqvist et al., 2017). One reason could be an overdispersion of the wind fields of the WRF ensemble. The stochastic perturbations were not calibrated, as a larger number of hindcasted events are necessary to be able to optimize the perturbations. Another issue is the roughness length of the sea surface, which is defined as a constant value in the applied WRF setup. Under severe storm conditions, the roughness of the sea surface should increase, resulting in a reduction of atmospheric kinetic energy and a corresponding limitation of wave growth. The storm events, Toini and Rafael, with the highest observed significant wave heights discussed by Björkqvist et al. (2017) were additionally hindcasted and are presented in the supplemental material. They seem to be less sensitive on the perturbations. Based on the short timeseries of observations, it is difficult to judge which significant wave height is still realistic.

A shortcoming of ~~this procedure is~~ the presented procedure can be the fact that the wave model runs were all started from the same initial state. This means that a certain time is needed until the different members diverge, especially as the total wave height is a combination of wind sea and swell. The later needs some time to travel, so that regions which are predominated by swell can be assumed to need a longer period to produce a reasonable spread with this setting. For the Baltic Sea, events with a strong influence of swell are infrequent (e.g. Broman et al., 2006; Soomere et al., 2012). French Guiana, for example, is one region which is swell dominated. Osinski et al. (2018) estimated the hundred-year return level of the significant wave height of northerly swell events at the French Guiana coast. Such events are generated in the Northern Atlantic and travel until the north-eastern coast of South America. For hindcasting such events with the demonstrated procedure, a large domain would be necessary and long lasting forecast horizons, so that the waves are already perturbed where generated and over their lifetime as well. This can lead to stronger deviations from the real past state.

3.3 Robustness of the ensemble spread depending on the ensemble size

Each ensemble member is a random draw of the PDF of the forecast / hindcast uncertainty. In the extreme case of having only two members, it is very unlikely that the most extreme cases are represented. By increasing the ensemble size, the probability is getting higher that the full range of uncertainty is sampled. For local area models, operational weather forecast centres produce ensembles with around ten to twenty members. At first sight, this number seems to be comparable to the presented study. If driving the regional model from a global ensemble with about 30 to 50 ensemble members, one can use a clustering technique to identify the most representative members instead of randomly selecting a small subsample, what improves the ensemble performance (e.g. Nuissier et al., 2012). If the ensemble is initialised several times per day, the different runs can be combined using the LAF approach (e.g. Raynaud and Bouttier, 2017). To predict the probability of the exceedance of a certain threshold, one can apply also neighbourhood techniques (e.g. Theis et al., 2005) or apply post-processing techniques like Bayesian Model Averaging (Raftery et al., 2005). Neither the initial and lateral boundary conditions come from a large ensemble in this study nor the application of neighbourhood or other post-processing techniques helps, because the ensemble members are used to

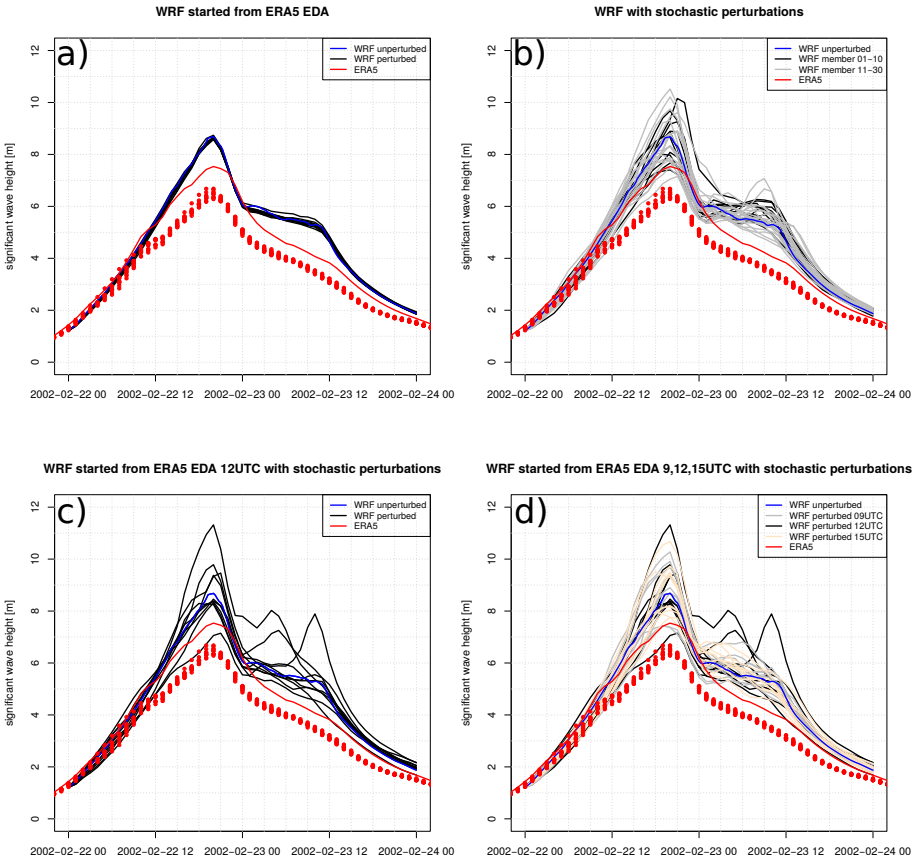

Figure 6. Significant wave height [m] at station Northern Baltic (21.00°E, 59.25°N, see ~~figure~~ Figure 1); driven by WRF ensemble based on a) initialisations at ~~2002-02-21 12UTC~~ 21 February 2002 12 UTC from ERA5 EDA, b) ERA5 reanalysis and stochastic perturbations, c) ERA5 EDA with SKEB and SPPT, and d) ERA5 EDA with SKEB and SPPT initialed additionally on same day at 09 and 15 UTC; ERA5 significant wave heigths from the ECWAM model in 0.36deg resolution; Shown ~~2002-02-22 21UTC~~22 February 2002 21 UTC.

drive a wave model. To get an idea how many ensemble members are reasonable in this case, 500 members have been generated with stochastic perturbations. From these 500 members, an ensemble with N members is generated, with N starting at 10 going until 300. Ten million samples of each ensemble of size N are selected by randomly choosing N out of the 500 members. The standard deviation is used as a measure for the ensemble spread and is calculated for each of the ten million samples of the ensemble of size N. Selecting N out of 500 members, ~~500 choose N~~ $\binom{500}{N}$ possibilities exist. This number ~~is largely above~~ exceeds ten millions for all tested ensemble sizes between 10 and 300. If the ensemble size is reasonable to get a robust estimate of the uncertainty, the spread should be relatively similar between each of the samples. ~~Figure~~

Figure 8 shows box-whisker plots for the ten million samples for ensemble sizes between 10 and 300 members.

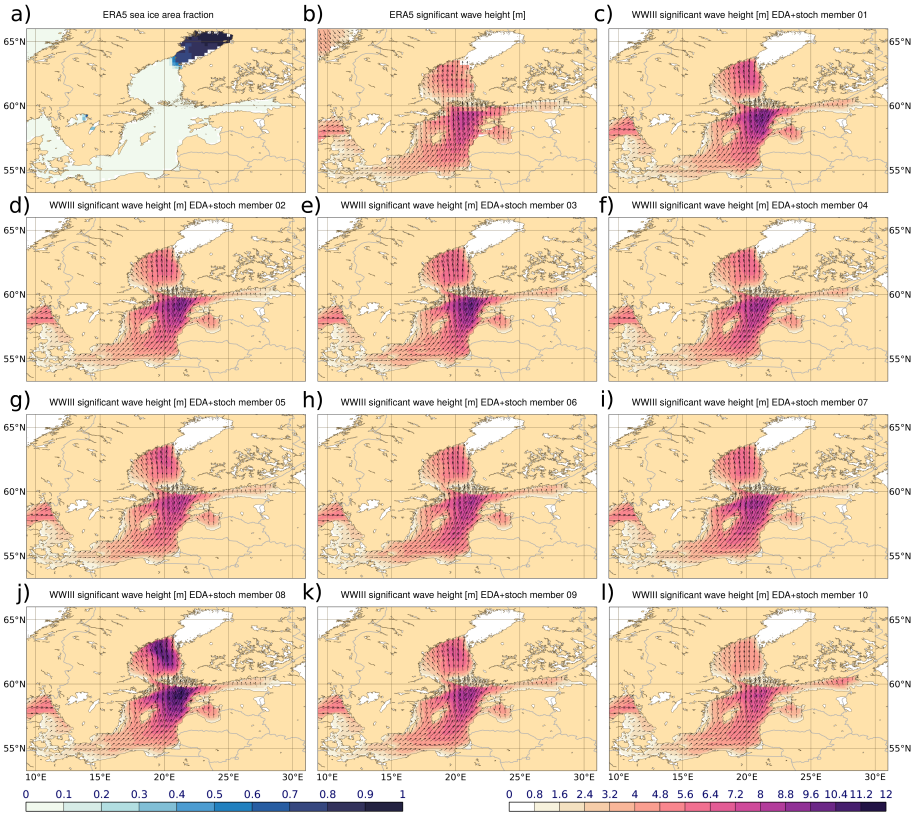

Figure 7. a) ERA5 sea ice area fraction [0-1]; b) ERA5 ECWAM significant wave height [m] with direction in meteorological convention and c-l) ten ~~Wavewatch III®~~ WAVEWATCH III® members driven by WRF ensemble initialised from ten ERA5 EDA members at ~~2002-02-21 12UTC~~ 21 February 2002 12 UTC with SPPT and SKEB. Shown ~~2002-02-22 21UTC~~22 February 2002 21 UTC.

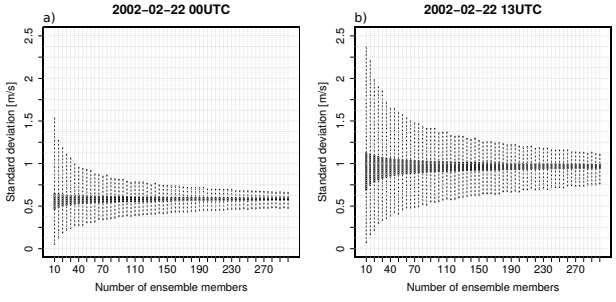

Figure 8. Box-whisker plots of the standard deviation of the 10 m wind speed at 19.39°E, 56.17°N of ten million samples of ensembles of size 10 to 300 randomly sampled from an ensemble with 500 members generated with WRF by applying SKEB and SPPT; shown a) 22 February 2002 00 and b) 13 UTC. Compare Fig. 2.

~~Box-whisker plots of the standard deviation of the 10 m wind speed at 19.39°E, 56.17°N of ten million samples of ensembles of size 10 to 300 randomly sampled from an ensemble with 500 members generated with WRF by applying SKEB and SPPT; shown 2002-02-22 00 and 13UTC. Compare Fig. 2.~~

The variation of the spread in the ten million samples twelve hours after initialisation is demonstrated in the left panel a). As it needs some time that the stochastic perturbations provoke spread between the ensemble members, there is a lead time dependence in the spread. The right panel b) presents a situation 25 hours after initialisation. At this time, the wind speed is very high, see Figure 2. In extreme situations, in which we are especially interested, we expect an higher uncertainty. This higher uncertainty is represented by a larger spread. All the ensembles with sizes between fifteen and hundred members show a median of the spread around one, at ~~February 22nd,~~ 22 February 2019 13 UTC. The ten member ensemble has a slightly lower median. With a higher uncertainty, a larger number of ensemble members is necessary to sample the entire uncertainty range. ~~By increasing the~~ With increasing ensemble size, ~~the estimation of the uncertainty~~ it is getting more ~~robust,~~ probable that the entire uncertainty range is sampled. This is why the range of the box-whisker plots is decreasing with increasing ensemble size. At ~~February 22nd,~~ 22 February 2019 00 UTC, the uncertainty is lower and / or the spread as a measure of uncertainty is not yet fully ~~developped~~ developed after twelve hours~~, why~~ . As the robustness of the ensemble spread seems to be dependent on the uncertainty, the range of the box-whisker plots is much inferior at 22 February 2019 00 UTC than thirteen hours later. To achieve a general statement about the ensemble size / spread relation, a much larger sample over a longer period must be investigated, but it can already be concluded that an ensemble size of only ten randomly generated members, as demonstrated in this application, can lead to a significant over- or underestimation of the uncertainty. ~~A~~ Depending on the application, the ensemble size needs to be selected by a compromise between the robustness of the uncertainty estimate and the computational ~~effort has to be found~~cost.

3.4 Impact of the spatiotemporal resolution of the atmospheric forcing ~~onto~~ on the significant wave height~~The~~

The numerical time step of a ~~high resolution ocean or wave model is normally below one~~ wave model can be counted in seconds (typical for explicit numerical schemes) or minutes (typical for implicit numerical schemes). The wave model therefore needs updated wind information e.g. every 30 seconds. This is done by interpolation from the wind forcing that is provided e.g. every hour or every third hour. A higher temporal resolution of atmospheric forcing data than one hour is normally not available~~, but atmospheric forcing fields are needed for all model timesteps. These can be produced for example by linear interpolation in time.~~ . If a variable in the ocean-/wave model to be driven has a short response time (e.g. surface current generated by wind compared to SST whose response is slower), and the variability of the atmospheric forcing in between the temporal resolution of the forcing fields is high, the result can be an under- or overestimation and a erroneous time evolution. One imaginable solution is to use maximum values during the output time interval of the atmospheric model, but this can lead to spatially inconsistent fields, especially if the time interval is very long. To test the impact of different temporal resolutions on the significant wave height, wind fields in 5 minutes resolution were produced with the 0.063~~°~~ deg setup. Figure 9 shows the wind field in 5, 15, 30 and 60 minutes resolution at one specific grid cell and the resulting significant wave height at the same location and time. It can be seen that the wind speed maximum in the 60 minutes resolution is about 0.25 m/s below the maxima of

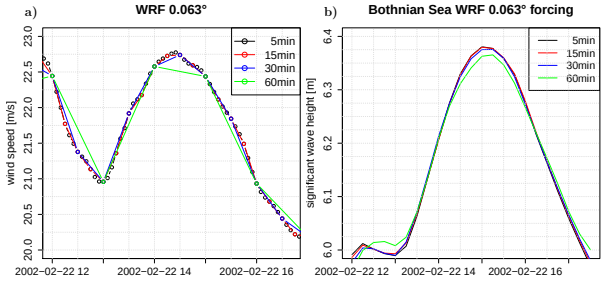

Figure 9. a) Wind speed [m/s] in 0.063~~°~~ deg setup and b) significant wave height [m] at 20.23°E and 61.8°N ; Testcase without sea ice

the higher temporal resolutions. ~~The~~ Between the higher temporal resolutions ~~do not differ so much~~of the wind data, the wind speed maxima are very close. The effect between the 60 minutes ~~and higher resolved temporal forcing~~ temporal resolution and a forcing in higher temporal resolution on the significant wave height is relatively low with about 2 cm. Systematic differences cannot be found based on the small sample, but ~~it~~ this sensitivity test indicates that the choice of the 15 minutes resolution is a reasonable compromise between a good representation of the extreme values and file size.

A stronger impact can be expected from the spatial resolution of the driving wind fields, because a coarser resolution of the atmospheric model can be assumed to produce lower extreme wind speeds as a grid cell represents the average value over the area it covers. By adapting, for example, the parameter betamax which describes the maximum value of wind-wave coupling, this difference can be compensated for. A better representation of the complex ~~orography of the coastlines in the~~ coastline of the Baltic sea as well as of the various islands is given by the higher resolved WRF model. For this reason, a difference in the spatial pattern ~~can be assumed~~of the significant wave height can be expected. A test with the coarsest (0.252deg) and the highest ~~resolutions~~ resolution (0.063deg) produced in this study has been conducted. The same parameter sets were used, as a calibration is not possible based on the short period hindcasted with WRF. Figure 10 shows the difference between these two forcings on one timestep in the significant wave height. One grid cell of the coarser WRF setup contains 16 grid cells of the higher resolved setup. Maxima as well as minima were found more extreme in the higher resolution with a higher spatial variability, what explains the higher but also the lower wave heights. It would be interesting to determine the remaining difference in the wave parameters provoked by the atmospheric forcings with different resolutions after a calibration of the wave model, done by applying an automatic and objective calibration procedure like for example the one proposed by Gorman and Oliver (2018). Tuomi et al. (2014) studied the effect of different spatiotemporal resolutions of the wind forcing on a wave model with a higher spatial resolution than applied here. A wave model with a higher resolution might benefit more from a higher resolution of the wind forcing.

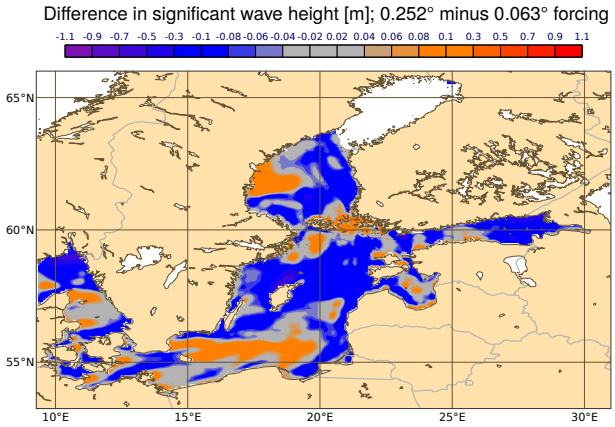

Figure 10. Difference in the significant wave height [m] in WAVEWATCH III® between simulation driven by WRF with 0.252~~°~~ deg and by 0.063~~°~~ deg in 15 minutes temporal resolution at ~~2002-02-22 21UTC~~22 February 2002 21 UTC. Part of the Gulf of Bothnia contains sea ice cover fractions above 0.5 why there are no waves.

4 Conclusions

Different approaches for hindcasting a single storm event in the Baltic Sea, which provoked a severe storm surge in February 2002, were tested in this study to create an ensemble hindcast of atmospheric (wind) and wave conditions based on a state-of-the-art atmospheric mesoscale model and a third generation wave model. The objective of the ensemble approach is a quantification of the uncertainty of the hindcast. The wave model was calibrated based on a publicly available regional reanalysis, and than validated with this dataset and also with forcing data produced with the atmospheric setup used in this study, as demonstrated in the supplemental material. A lagged-average WRF forecast ensemble showed only little spread with initial and lateral boundary conditions based entirely on the high resolution ERA5 reanalysis fields. The spread of the LAF ensemble can not be easily adapted.[14]

A domain shifting approach with ERA5, in which the input fields are shifted into all directions by one grid cell, shows a more or less comparable spread to the LAF ensemble, with the same advantage of using the high resolution reanalysis data only. The number of ensemble members and the spread of the ensemble is limited to the number of reasonable shifts. Too large shifts can be expected to degrade the hindcast.[15]

[14] A weighting of the different realisations (in this case of the wave model) by giving the runs more weight which are expected to have a lower error is possible. In this case more weight to the runs should be given which have smaller errors in the verification of a large sample of hindcasts (e.g. more weight to runs with lower lead time to the desired event), but this can be expected to not strongly enlarge the ensemble spread.

[15] As the WRF model has a finer resolution, shifts different than multiples of one grid cell by adding or substracting an offset onto the coordinates of the ERA5 grid would change the interpolation for the WRF initial and lateral boundary conditions. This was not tested, and it was also not investigated systematically if members generated from smaller shifts are closer to the unpertubed run or if shifts into a certain direction (e.g. into flow direction or perpendicular to it) lead to different spreads than shifts into other directions, which would mean that there are systematic differences between the members to be taken into account by the interpretation of the ensemble data. A test with shifts of two and three grid cells into north, west, south and east direction were tested and indicate that there are systematic differences.

Starting WRF from the ERA5 EDA members show also a comparable spread as with the two other approaches. The disadvantage is the coarser resolution of the initial fields. Fine ~~scales~~ scale structures are not present in these fields so that the ensemble spread is limited~~by this fact~~. Stochastic perturbations produce a much larger spread, but need some time to develop. ~~As the~~ The first twelve hours are not used ~~, because of the~~ in this study, because they are assumed to be affected by a model spin-up~~, this is not really a shortcoming .~~. This is not a shortcoming for a hindcast procedure.

A combination of stochastic ~~pertubations~~ perturbations and an initialisation from the ERA5 EDA fields produces also deviations from the unperturbed runs which are not present by only using the stochastic perturbations. This approach is especially interesting and is close to what is used in meterological weather forecast centres for the operational forecasts. The wind fields from this ensemble hindcast produce also a large spread in the wave model. A visual comparison with the ERA5 wave model ensemble of data assimilations indicates that this spread is more reasonable than using the first three discussed ensemble generation approaches. The peak of the significant wave height in the Baltic proper of the most extreme members is, however, with about 11 m strongly exceeding existing observations in this region. One possible reason could be an overdispersion of the ensemble system. Another important factor is the roughness of the sea surface and its impact on the dynamics of the storm. In the presented setup, the roughness length of the sea surface is defined as a constant value. A coupling of the atmospheric with the wave model would allow to adapt the roughness length depending on the sea surfcace conditions and can lead to a limitation of the wave growth.

The robustness of the spread depending on the ensemble size was tested by randomly generating ensembles with different sizes (10 to 300 members) from an ensemble hindcast with 500 members. For small ensembles, the range of the ensemble spread can differ largely depending which members were randomly selected. In operational services, this problem is tackled by selecting for example already representative members from a larger global ensemble. To achieve a comparable robust estimate of the uncertainty, the ensemble size for the here presented approach without pre-selection of ensemble members must be larger than the one of operational local area model ensembles.

Another source of uncertainty arises from the spatiotemporal discretisation of the ~~atmopsheric~~ atmospheric model and the resulting forcing fields for the wave model. Errors introduced by a coarse temporal resolution of the driving wind fields in the significant wave height are relatively small in this event testcase. For a strong event with a significant wave height of about 6.3 m, the difference between a 5 and 60 minutes temporally resolved wind forcing is only on the order of 2 cm. Between 15 minutes and 5 minutes temporal resolution, the impact on the wave height is negligible for the demonstrated case. The horizontal resolution has a much stronger impact. This can be potentially be corrected by calibrating the model to the different wind forcings. It would be interesting to estimate the remaining difference, but this was not ~~possbile~~ possible in the framework of this study as a calibration of the models is not feasible based on a hindcast of only a single event.

A combination of ERA5 EDA fields as initial conditions and the stochastic perturbations showed the ability to produce a larger spread than with the other demonstrated approaches. Stochastic perturbations haven't been tuned in this study. Producing longer timeseries, for tuning and validating the model could lead to a reasonable measure of the hindcast uncertainty on the regional scale~~and~~ . Operational atmospheric and wave products exist with a comparable or even higher resolution than applied here, whose quality is superior to what can be reached with the demonstrated procedure as they include state-of-the-art data

assimilation techniques. The application of such operational products is however limited by the available periods and also by the inhomogeneity of the datasets. The demonstrated approach allows to adapt the spatiotemporal resolution and ~~number of members to the applications needs~~the ensemble size to specific research questions for event based hindcasts in a homogeneous manner over the entire available ERA5 period.

Code availability. The WRF source code is available from https://github.com/wrf-model/WRF/releases and the WAVEWATCH III® from https://github.com/NOAA-EMC/WW3

Data availability. ERA5 and the UERRA/Harmonie-v1 reanalysis can be retrieved from the Climate data store at https://cds.climate.copernicus.eu.

Sample availability. Ensemble hindcasts of wind and wave fields presented in this study can be requeested by contacting the corresponding author.

Author contributions. RDO is responsible for the concept of this study, prepared the configurations of WRF and WAVEWATCH III®, conducted the simulations and prepared the presentation of the results. HR was involved in the discussion of the results and in the preparation of the manuscript.

Competing interests. The authors declare that there is no conflict of interest.

Acknowledgements. This study was financed by the Bonus Micropoll project, which has received funding from BONUS (Art 185), funded jointly by the EU and Baltic Sea national funding institutions. For the calibration and validation of the Baltic Sea ~~Wavewatch~~ WAVEWATCH III® setup, computing resources at the HLRN were consumed and E.U. Copernicus Marine Service Information were used. The simulations in this study were generated using Copernicus Climate Change Service Information (2018/2019). The research and work leading to the UERRA data set used in this study has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement № 607193. We would like to thank the WRF and ~~Wavewatch~~ WAVEWATCH III® developers for providing their models over Github.

References

Ardhuin, F., H C Herbers, T., O'Reilly, W., and Jessen, P.: Swell Transformation across the Continental Shelf. Part I: Attenuation and Directional Broadening, Journal of Physical Oceanography, 33, 1921, https://doi.org/10.1175/1520-0485(2003)033<1921:STATCS>2.0.CO;2, 2003.

Ardhuin, F., Rogers, E., Babanin, A. V., Filipot, J.-F., Magne, R., Roland, A., van der Westhuysen, A., Queffeulou, P., Lefevre, J.-M., Aouf, L., and Collard, F.: Semiempirical Dissipation Source Functions for Ocean Waves. Part I: Definition, Calibration, and Validation, Journal of Physical Oceanography, 40, 1917–1941, https://doi.org/10.1175/2010JPO4324.1, https://doi.org/10.1175/2010JPO4324.1, 2010.

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, Natural Hazards and Earth System Sciences, 17, 1653–1658, https://doi.org/10.5194/nhess-17-1653-2017, https://www.nat-hazards-earth-syst-sci.net/17/1653/2017/, 2017.

Bouttier, F., Vié, B., Nuissier, O., and Raynaud, L.: Impact of Stochastic Physics in a Convection-Permitting Ensemble, Monthly Weather Review, 140, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1, https://doi.org/10.1175/MWR-D-12-00031.1, 2012.

Brankart, J.-M., Candille, G., Garnier, F., Calone, C., Melet, A., Bouttier, P.-A., Brasseur, P., and Verron, J.: A generic approach to explicit simulation of uncertainty in the NEMO ocean model, Geoscientific Model Development, 8, 1285–1297, https://doi.org/10.5194/gmd-8-1285-2015, https://www.geosci-model-dev.net/8/1285/2015/, 2015.

Broman, B., Hammarklint, T., Kalev, R., Soomere, T., and Valdmann, A.: Trends and extremes of wave fields in the north-eastern part of the Baltic Prope, Oceanologia, 48, 2006.

Buizza, R.: Impact of Horizontal Diffusion on T21, T42, and T63 Singular Vectors, Journal of the Atmospheric Sciences, 55, 1069–1083, https://doi.org/10.1175/1520-0469(1998)055<1069:IOHDOT>2.0.CO;2, https://doi.org/10.1175/1520-0469(1998)055<1069:IOHDOT>2.0.CO;2, 1998.

Buizza, R., Milleer, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 125, 2887–2908, https://doi.org/10.1002/qj.49712556006, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712556006, 1999.

Buizza, R., Leutbecher, M., and Isaksen, L.: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System, Quarterly Journal of the Royal Meteorological Society, 134, 2051–2066, https://doi.org/10.1002/qj.346, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.346, 2008.

Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), accessed on 26th march 2019, https://cds.climate.copernicus.eu/cdsapp#!/home, 2017.

ECMWF: Part VII: ECMWF Wave Model, no. 7 in IFS Documentation, ECMWF, https://www.ecmwf.int/node/16651, 2016.

Farina, L.: On ensemble prediction of ocean waves, Tellus A, 54, 148–158, https://doi.org/10.1034/j.1600-0870.2002.01301.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-0870.2002.01301.x, 2002.

Gorman, R. M. and Oliver, H. J.: Automated model optimisation using the Cylc workflow engine (Cyclops v1.0), Geoscientific Model Development, 11, 2153–2173, https://doi.org/10.5194/gmd-11-2153-2018, https://www.geosci-model-dev.net/11/2153/2018/, 2018.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept, Tellus A, 57, 219–233, https://doi.org/10.1111/j.1600-0870.2005.00103.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.2005.00103.x, 2005.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Monthly Weather Review, 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hoffman, R. N. and Kalnay, E.: Lagged average forecasting, an alternative to Monte Carlo forecasting, Tellus A, 35A, 100–118, https://doi.org/10.1111/j.1600-0870.1983.tb00189.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.1983.tb00189.x, 1983.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley & Sons Ltd.: Chichester, UK., 2003.

Leutbecher, M., Lock, S.-J., Ollinahob, P., Lang, S. T. K., Balsamo, G., Bechtold, P., Bonavita, M., Christensenc, H. M., Diamantakis, M., Dutra, E., English, S., Fisher, M., Forbes, R. M., Goddard, J., Haiden, T., Hogan, R. J., Jurickec, S., Lawrence, H., MacLeodc, D., Magnusson, L., Malardel, S., Massart, S., Sandu, I., Smolarkiewicz, P. K., Subramanianc, A., Vitart, F., Wedi, N., and Weisheimer, A.: Stochastic representations of modeluncertainties at ECMWF:State of the art and future vision, Tech. rep., European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, England, 2016.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H. M., Diamantakis, M., Dutra, E., English, S., Fisher, M., Forbes, R. M., Goddard, J., Haiden, T., Hogan, R. J., Juricke, S., Lawrence, H., MacLeod, D., Magnusson, L., Malardel, S., Massart, S., Sandu, I., Smolarkiewicz, P. K., Subramanian, A., Vitart, F., Wedi, N., and Weisheimer, A.: Stochastic representations of model uncertainties at ECMWF: state of the art and future vision, Quarterly Journal of the Royal Meteorological Society, 143, 2315–2339, https://doi.org/10.1002/qj.3094, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3094, 2017.

Lorenz, E. N.: Deterministic Nonperiodic Flow, Journal of the Atmospheric Sciences, 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963.

Murphy, A. H. and Winkler, R. L.: Diagnostic verification of probability forecasts, International Journal of Forecasting, 7, 435–455, https://EconPapers.repec.org/RePEc:eee:intfor:v:7:y:1992:i:4:p:435-455, 1992.

Nuissier, O., Joly, B., Vié, B., and Ducrocq, V.: Uncertainty of lateral boundary conditions in a convection-permitting ensemble: a strategy of selection for Mediterranean heavy precipitation events, Natural Hazards and Earth System Sciences, 12, 2993–3011, https://doi.org/10.5194/nhess-12-2993-2012, https://www.nat-hazards-earth-syst-sci.net/12/2993/2012/, 2012.

Ollinaho, P., Lock, S.-J., Leutbecher, M., Bechtold, P., Beljaars, A., Bozzo, A., Forbes, R. M., Haiden, T., Hogan, R. J., and Sandu, I.: Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble, Quarterly Journal of the Royal Meteorological Society, 143, 408–422, https://doi.org/10.1002/qj.2931, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2931, 2017.

Osinski, R., Lorenz, P., Kruschke, T., Voigt, M., Ulbrich, U., Leckebusch, G. C., Faust, E., Hofherr, T., and Majewski, D.: An approach to build an event set of European windstorms based on ECMWF EPS, Natural Hazards and Earth System Sciences, 16, 255–268, https://doi.org/10.5194/nhess-16-255-2016, https://www.nat-hazards-earth-syst-sci.net/16/255/2016/, 2016.

Osinski, R., Dalphinet, A., Aouf, L., and Palany, P.: Estimation of the hundred year return level of the significant wave height for the French Guiana coast, Brazilian Journal of Oceanography, 66, 325 – 334, http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1679-87592018000400325&nrm=iso, 2018.

Pardowitz, T., Befort, D. J., Leckebusch, G. C., and Ulbrich, U.: Estimating uncertainties from high resolution simulations of extreme wind storms and consequences for impacts, Meteorologische Zeitschrift, 25, 531–541, https://doi.org/10.1127/metz/2016/0582, http://dx.doi.org/10.1127/metz/2016/0582, 2016.

Peckham, S. E., Smirnova, T. G., Benjamin, S. G., Brown, J. M., and Kenyon, J. S.: Implementation of a Digital Filter Initialization in the WRF Model and Its Application in the Rapid Refresh, Monthly Weather Review, 144, 99–106, https://doi.org/10.1175/MWR-D-15-0219.1, https://doi.org/10.1175/MWR-D-15-0219.1, 2016.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133, 1155–1174, https://doi.org/10.1175/MWR2906.1, https://doi.org/10.1175/MWR2906.1, 2005.

Raynaud, L. and Bouttier, F.: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts, Quarterly Journal of the Royal Meteorological Society, 143, 3037–3047, https://doi.org/10.1002/qj.3159, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3159, 2017.

Ricchi, A., Miglietta, M. M., Bonaldo, D., Cioni, G., Rizza, U., and Carniel, S.: Multi-Physics Ensemble versus Atmosphere–Ocean Coupled Model Simulations for a Tropical-Like Cyclone in the Mediterranean Sea, Atmosphere, 10, https://doi.org/10.3390/atmos10040202, http://www.mdpi.com/2073-4433/10/4/202, 2019.

Ridal, M., Olsson, E., Unden, P., Zimmermann, K., and Ohlsson, A.: Uncertainties in Ensembles of Regional Re-Analyses - Deliverable D2.7 HARMONIE reanalysis report of results and dataset, http://www.uerra.eu/component/dpattachments/?task=attachment.download&id=296, 2017.

Seifert, T., Tauber, F., and Kayser, B.: A high resolution spherical grid topography of the Baltic Sea - 2nd edition, in: Baltic Sea Science Congress, Stockholm 25-29, Poster No. 147, www.io-warnemuende.de/iowtopo, 2001.

Shutts, G.: A kinetic energy backscatter algorithm for use in ensemble prediction systems, Quarterly Journal of the Royal Meteorological Society, 131, 3079–3102, https://doi.org/10.1256/qj.04.106, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.04.106, 2005.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D., and yu Huang, X.: A Description of the Advanced Research WRF Model Version 4, Tech. rep., NCAR/TN-556+STR, https://doi.org/doi:10.5065/1dfh-6p97, 2019.

Soomere, T.: Wind wave statistics in Tallinn Bay, Boreal Environment Research, 10, 103–118, 2005.

Soomere, T., Behrens, A., Tuomi, L., and Nielsen, J. W.: Wave conditions in the Baltic Proper and in the Gulf of Finland during windstorm Gudrun, Natural Hazards and Earth System Sciences, 8, 37–46, https://doi.org/10.5194/nhess-8-37-2008, https://www.nat-hazards-earth-syst-sci.net/8/37/2008/, 2008.

Soomere, T., Weisse, R., and Behrens, A.: Wave climate in the Arkona Basin, the Baltic Sea, Ocean Science, 8, 287–300, https://doi.org/10.5194/os-8-287-2012, https://www.ocean-sci.net/8/287/2012/, 2012.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J. H., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H. A., Wilson, L., and Yamaguchi, M.: The TIGGE Project and Its Achievements, Bulletin of the American Meteorological Society, 97, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1, https://doi.org/10.1175/BAMS-D-13-00191.1, 2016.

The WAVEWATCH III® Development Group (WW3DG): User manual and system documentation of WAVEWATCH III® version 6.07. Tech. Note 333, 2019.

Theis, S. E., Hense, A., and Damrath, U.: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach, Meteorological Applications, 12, 257–268, https://doi.org/10.1017/S1350482705001763, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482705001763, 2005.

Tolman, H. L.: A Third-Generation Model for Wind Waves on Slowly Varying, Unsteady, and Inhomogeneous Depths and Currents, Journal of Physical Oceanography, 21, 782–797, https://doi.org/10.1175/1520-0485(1991)021<0782:ATGMFW>2.0.CO;2, https://doi.org/10.1175/1520-0485(1991)021<0782:ATGMFW>2.0.CO;2, 1991.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NCEP and the Breeding Method, Monthly Weather Review, 125, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2, 1997.

Tuomi, L., Kahma, K., and Pettersson, H.: Wave hindcast statistics in the seasonally ice-covered Baltic Sea, Boreal Environment Research, 16, 451–472, 2011.

Tuomi, L., Pettersson, H., Fortelius, C., Tikka, K., Björkqvist, J.-V., and Kahma, K. K.: Wave modelling in archipelagos, Coastal Engineering, 83, 205–220, https://doi.org/https://doi.org/10.1016/j.coastaleng.2013.10.011, http://www.sciencedirect.com/science/article/pii/S0378383913001671, 2014.

Wang, W., Bruyère, C., Duda, M., Dudhia, J., Gill, D., Kavulich, M., Werner, K., Chen, M., Lin, H.-C., Michalakes, J., Rizvi, S., Zhang, X., Berner, J., Munoz-Esparza, D., Reen, B., and Fossel, S. H. K.: User's Guide for the Advanced Research WRF (ARW) Modeling System, Version 4, http://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V4/WRFUsersGuide.pdf, 2019.

Yildirim, B. and Karniadakis, G. E.: Stochastic simulations of ocean waves: An uncertainty quantification study, Ocean Modelling, 86, 15 – 35, https://doi.org/https://doi.org/10.1016/j.ocemod.2014.12.001, http://www.sciencedirect.com/science/article/pii/S1463500314001759, 2015.