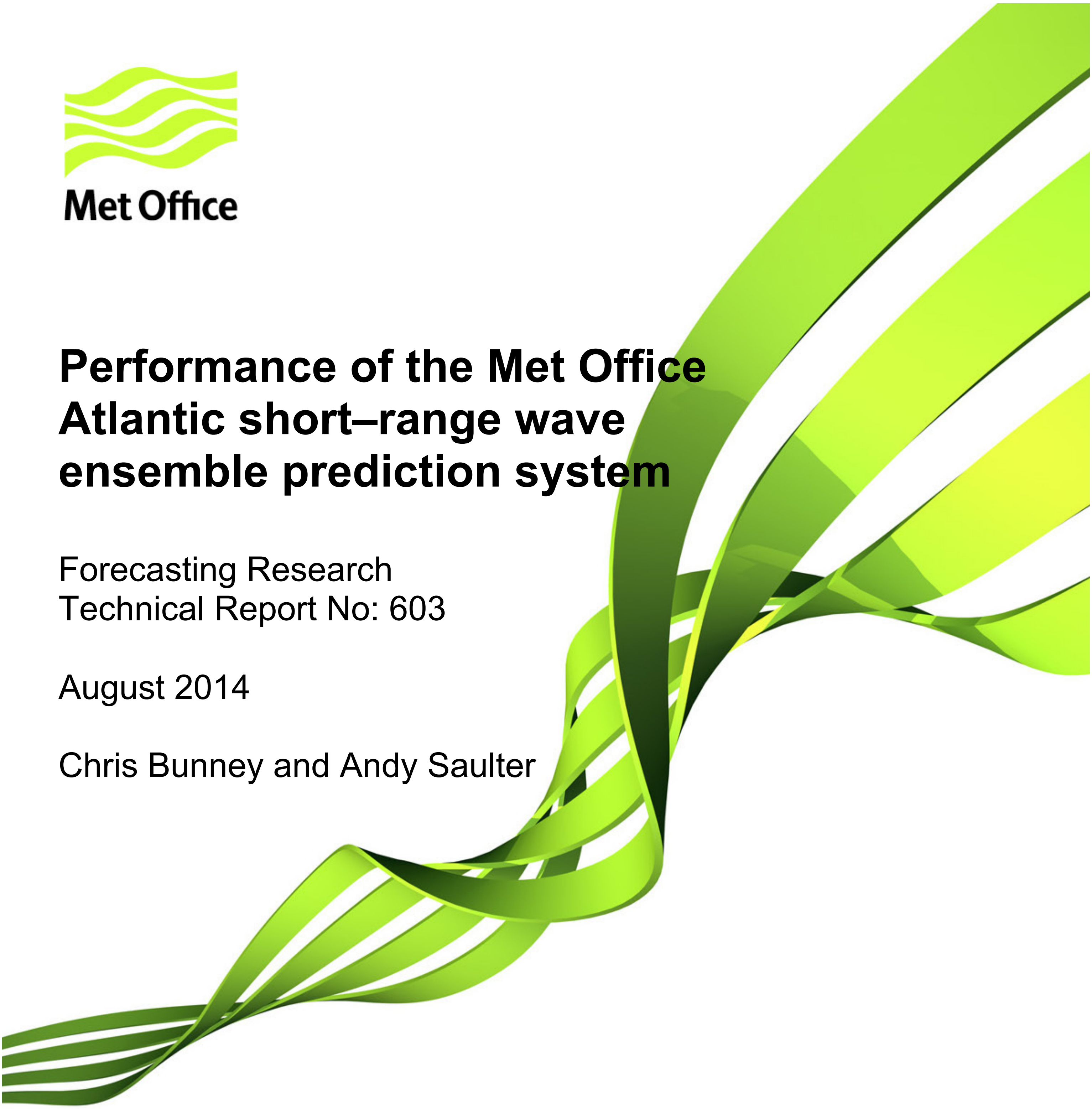
Performance of the Met Office Atlantic short–range wave ensemble prediction system

Forecasting Research
Technical Report No: 603

August 2014

Chris Bunney and Andy Saulter



Contents

Executive Summary

This is a summary of the key results from the verification of the Met Office's trial Atlantic wave ensemble system concentrating on two areas around the U.K. The Atlantic ensemble system is anticipated to be a working model for U.K. wide 1-7 day probabilistic wave forecasts. Currently, the system runs out to 3 days and this document focuses on probabilistic verification at that lead time; this lead time happens to sit on the cusp for predictability of synoptic scale storm systems and is therefore relevant for assessing the short-range performance that is important for the operational decision maker.

The model has been assessed in terms of climate prediction, deterministic error and ensemble spread relationships, model member selection and probability of threshold exceedance. Verification includes a novel application of observational error data in providing an ideal performance baseline.

Conclusions for the performance in the two U.K. areas are:

- Overall good performance
- Systematic low bias at high wave heights in the North Sea – linked to current model physics not growing waves strongly enough in short fetch areas
  - Suggests move to WAM based source terms will be a sensible future enhancement.
- Spread is a good indicator of model skill
  - This is strongly correlated to the wave conditions.
- Some evidence of under-spread; however there is a significant contribution from observational errors and this is dealt with using an *ideal model performance* baseline derived from model data with a novel application of observation error estimator.
- Probability forecasts at this range are significantly influenced by local bias – this is an important issue for short range forecasts.

Introduction

This document summarises the key findings from verification of Met Office trial Atlantic wave ensemble prediction model, in terms of potential application for Environment Agency (EA) coastal flood forecasting.

The wave prediction systems have been run pre-operationally at the Met Office since June 2013, as part of the EU funded MyWave project. The Atlantic configuration is intended to be used as a working model for probabilistic predictions in the 1 – 7 day timescales. In its current pre-operational state the system has a forecast lead time of 3 days (which sits on the cusp for predictability of synoptic scale storm systems) and therefore the probabilistic verification focuses on that lead time. It is intended to extend the forecast length to 7 days in its final operational incarnation.

The Atlantic wave ensemble prediction system

For an in-depth description of the wave ensemble systems and MOGREPS (Met Office Global and Regional Ensemble Prediction System), see the MyWave-D3.1 report (Bunney et al., 2013). A short overview of the Met Office's wind and wave ensemble systems is provided here.

The Atlantic wave model is formulated on a multi-resolution Spherical Multi Cell grid (SMC; Li, 2012) which provides increasingly finer resolution grid cells towards the coastline from 25/12/6km (Figure 1). The forecast length is 72 hours and forcing is provided via 10m wind fields from the Global MOGREPS (Bower et al., 2008) configuration (MOGREPS-G) which has a spatial resolution of 40km. The Tolman & Chalikov (1996) wave physics package is used for wave growth and dissipation in the model, although a newer WAM4 (Bidlot et al., 2012) is currently being tested as a replacement. It is anticipated that the forecast length of the model will be increased to 7 days in early 2015.

The wave ensemble system is a "lagged ensemble" comprised of 24 members. 11 ensemble members plus a control member are run at each cycle; the full 24 member ensemble is made up by taking the remaining members from the previous T-6 run – hence the term "lagged" ensemble. This is the approach used by the driving MOGREPS models and hence must be adopted by the wave ensemble.

All perturbations to the wave model are provided via the surface wind fields generated by the MOGREPS system, i.e. the wave model does not include a perturbation scheme, but relies on spread being introduced from the upstream model. The MOGREPS system generates its initial perturbations using the Ensemble Transform Kalman Filter (ETKF) which are applied to a control analysis field to provide initial conditions for each member (Bishop et al., 2001; Bowler et al., 2008).

To maintain a memory of the spread in the swell fields within the wave ensemble (i.e. wave fields that have been generated via some remote storm and are now freely propagating), the restart files for each member are cycled between runs. This is a different approach to member initialisation compared to other forecasting centres (such as ECMWF) where each member starts from the same control restart file. The argument for starting each member from initial conditions generated by the control is that the spread in an ensemble system should be related to the error in the forecast. For an analysis at T0, the error should be low (especially if some form of wave data assimilation is used). However, the error at T0 is not zero therefore the spread in the system at T0 should also not be zero. As the Met Office wave ensemble system does not employ any data assimilation, the error at T0 will be comparable to the error at T+6, which is the forecast hour that the restart for the next cycle is generated at. Therefore, cycling the restart files for each member has the advantage of providing a small amount of spread at T+0 and maintaining a memory of the spread in the swell generated from the previous cycles.

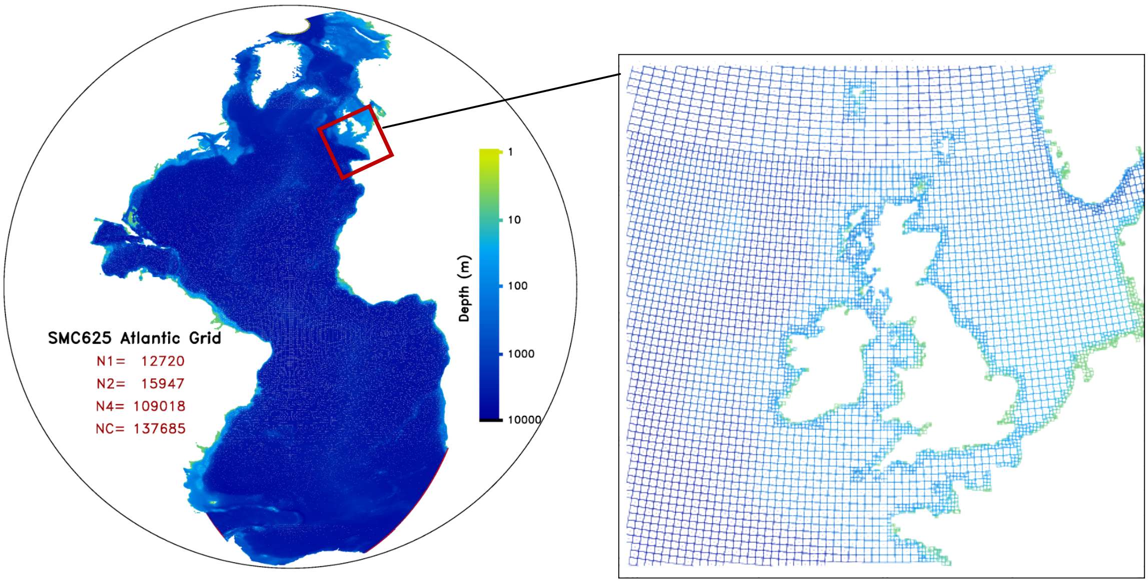


Figure 1 – Atlantic wave model domain (left) with detail of variable resolution grid cells around the coast of the U.K. and North West Europe. Grid cell sizes are 25, 12 and 6 km.

Verification procedure

An in-depth verification of the overall Atlantic domain is available as part of the "Atlantic-Euro zone" verification report produced for the MyWave project (Bunney, 2014). The report focuses on regional verification around the U.K. based on newly proposed EA requirements for use of ensemble data.

Observations

Two sets of observational data sources have been used for verification of the Atlantic and U.K. wave ensembles: in-situ buoy data and remotely sensed satellite data from the EUMETSAT JASON2 mission. For this report, the verification is focussed on the total significant wave height (H_s) and the 10 metre wind speed (U_{10}). Table 1 details the data collection periods and collocation methodology applied to the observation data sets.

Whilst the buoy data has a very good temporal resolution (data is usually hourly or half-hourly) the distribution of the buoys is mainly concentrated around the European and U.S. coastlines. It should be borne in mind that whilst the wave model does include shallow water physics to represent transformation of the waves as they propagate towards shallower water around the coast, the spatial resolution and lack of tidal elevations may introduce model errors at observing platforms in very shallow water (<20m depth).

The JASON2 satellite data has a much better spatial distribution (although the regular orbit track means that there are still unobserved areas) but the temporal coverage of any single point is limited to the 10 day return period of the satellite (OSTM/JASON2, 2011).

There is also the issue of land contamination causing large errors in the significant wave heights within ~15km of the coast. To avoid this impacting on the verification results, a 1x1 grid cell mask is applied to the model sea points adjacent to the coastline to remove these locations from the collocation.

Source	Period	Collocation method	Notes
Buoy	8 months: Aug ’13 – Mar ‘14	Nearest neighbour	Concentrated mostly around coastal areas. Good temporal resolution
JASON-2	6 months: Oct ’13 to Mar ‘14	Area average over model grid cell	Ku band *Hs*, altimeter derived U_{10}. Good spatial resolution, but sparser temporal coverage.

Table 1 - Observational data sets used in ensemble verification

Geographic Partitioning

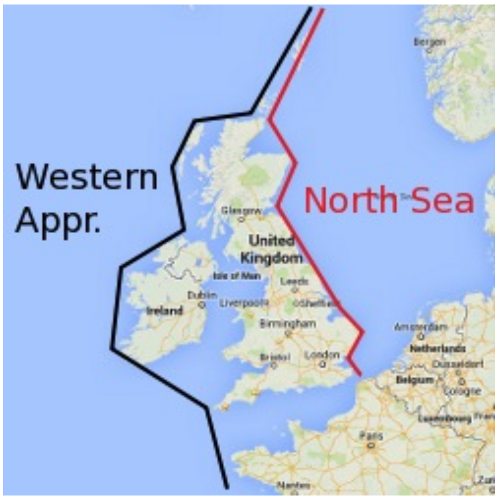


Figure 2 - Western Approaches and North Sea domains used for verification inter-comparison in the U.K. waters.

Although the Atlantic model coarse resolution in the open ocean, the refined-resolution grid cell scheme used means that the resolution around the U.K is considerable finer (approximately 6 km) which allows smaller scale coastal and bathymetric features to be resolved. The U.K. domain is an interesting area for model verification as it combines areas with varying wave climates. The Western Approaches to the U.K. [herein WA] are open to the Atlantic and hence receive mature, well developed wave systems generated by mid-Atlantic depressions. Conversely, the North Sea [herein NS], located between the east coast of U.K. and mainland Europe (see Figure 2) is sheltered from westerly Atlantic swell and can generally be considered a fetch limited wave generation area where waves tend to be much younger and developed by local winds. However, in strong northerly or southerly wind conditions, the fetch length of the North Sea is relatively long and it is not unusual to experience very large waves in the central North Sea.

Idealised Performance

When verifying any model system, there are two types of errors that must be taken into account; the model errors (which are usually the objective of the verification) and the observational errors (which are often unknown, but can be estimated). The existence of observational error makes achieving a perfectly performing model impossible unless it is accounted for. One method is to add a certain amount of “error noise” to the model predictions; an approach undertaken by Saetra et al. (2004b). An alternative approach that is adopted in this study is to produce a set of *pseudo-observations* that represent the *ideal* model performance. These pseudo-observations are generated by taking a random draw from the ensemble members over the entire forecast dataset being verified. To this, a heteroskedastic observational error estimate is added to each randomly selected value to generate a data set that the model can be verified against. The advantage of these pseudo-observations is that they give an idea of the *ideal model performance* where the *only* error is from the observational error estimate. Where appropriate, the idealised model performance has been presented on the probabilistic verification plots in the following sections. Full details of the generation and application of these pseudo-observations are detailed in the MyWave report D4.3 (Palmer & Saulter, 2013).

Replication of wave climate

A good starting point for any forecast model verification is to test whether it can reproduce the climatology of the observations. A quantile-quantile (Q-Q) plot is an often used metric for this kind of comparison and gives a measure of the model’s ability to produce the same distribution of wave heights in the domain as the observed distribution. As this metric compares distributions, rather than collocated model/observation pairs, it gives a measure of how well represented the various wave heights and wind speeds values are within the verification period. The x-axis shows the observed parameter quantiles and the y-axis the quantiles of modelled values. Ideally, the values for each quantile should be the same (i.e. lie on the diagonal line). A point below the diagonal suggests that the model is under-representing the values at that quanitlie, whilst a point above suggests model over-representation.

Figure 13 shows the Q-Q distribution plots for significant wave height versus combined in-situ buoy and JASON2 data. The model data is taken at forecast hours 1 – 6 as this is the period when the model errors should be the smallest. The left panels show results for locations in the WA and right panels for the NS. The top panels show the full Q-Q range from 1 – 99%iles in 1%ile increments. However, as it is often the quality of the higher, more exceptional, wave height forecasts that are of interest, the bottom panels show detail of the upper 90 – 99.9%iles.

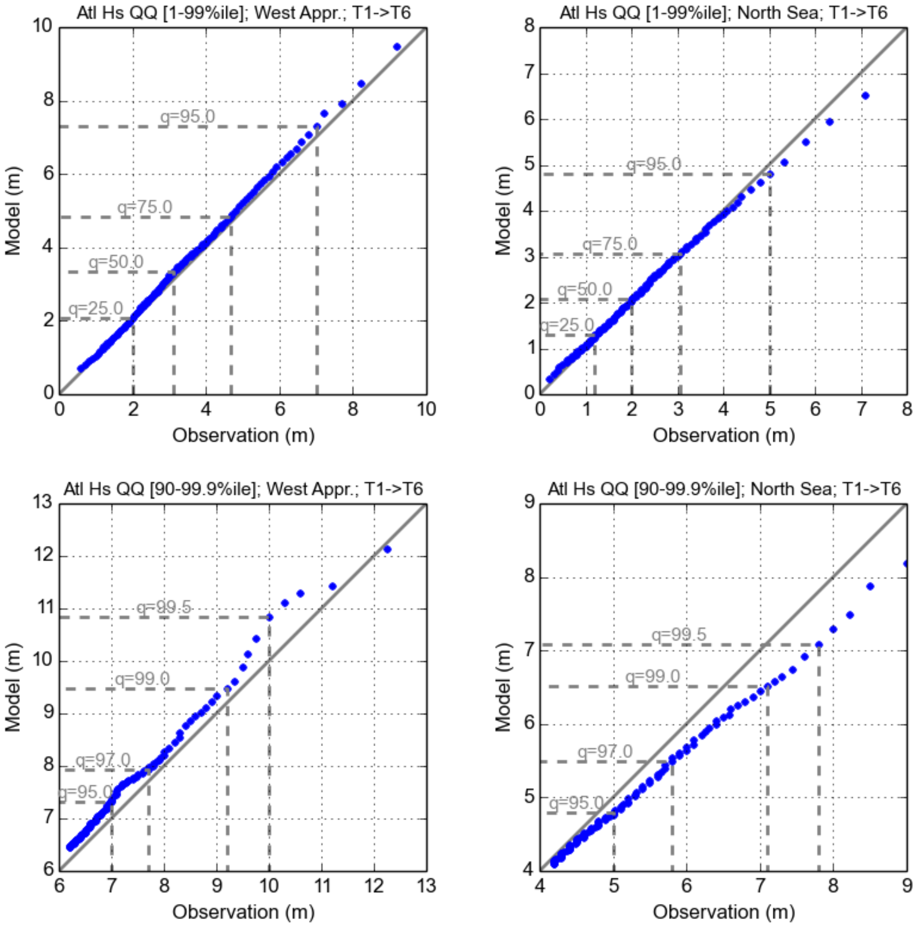


Figure 3 - QQ plots of model versus observations for significant wave height at T1-T6. Top panels: 1st - 99th percentile; bottom panels: 90.0th - 99.9th percentile. Left panels: Western Approaches; right: North Sea.

Figure 3 shows that both regions have good representation up to ~75%ile (equating to a H_s of 4.5m in the WA and 3m in the NS). However, there is a clear trend for the higher wave heights to be over-represented in the Western Approaches and under-represented in the North Sea. As errors in wave growth are very much influenced by errors in the driving wind fields, an initial check is made on the Q-Q plots for wind speed in both regions (Figure 4). The wind speed is very well represented in both regions up to about the 99%ile where both regions exhibit a slight under-representation of the strongest winds (> ~22m/s). This is a very different trend to the wave height distributions so it is unlikely that the deviation in wave height climate is a downstream impact of the wind speed climate representation.

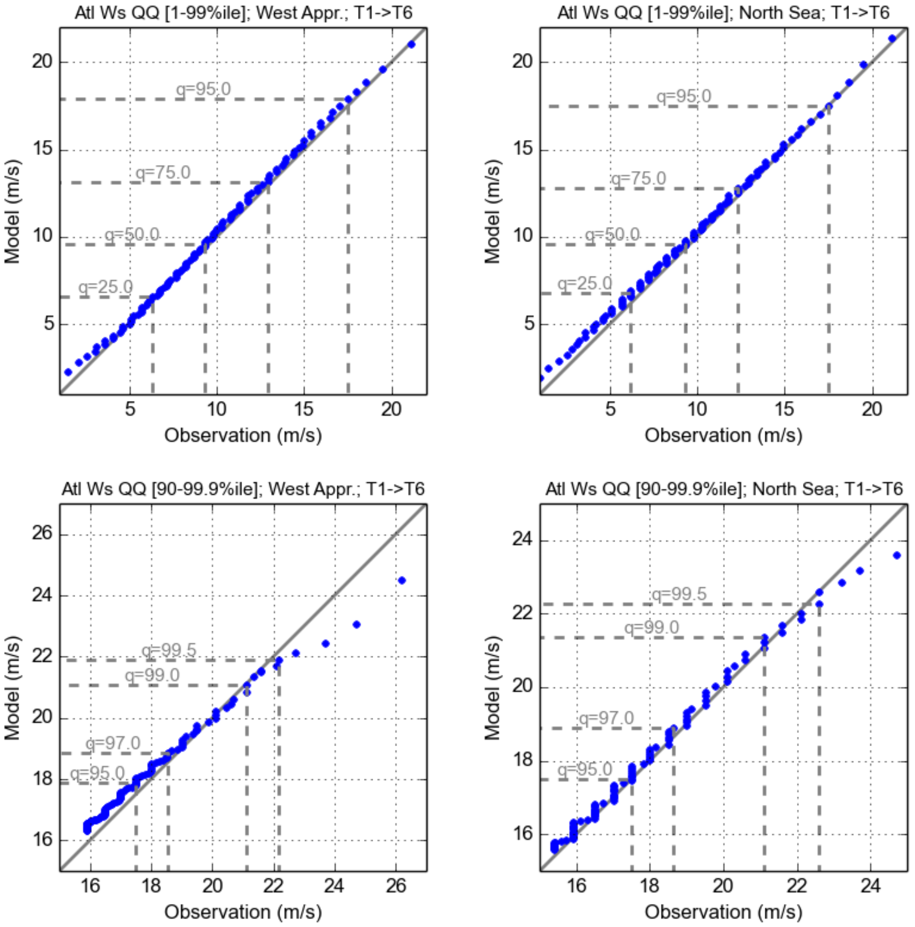


Figure 4 - QQ plots of model vs. observations for significant wave height at day 3. Top panels: 1st - 99th percentile; bottom panels: 99.0th - 99.9th percentile. Left panels: Western Approaches; right: North Sea.

The opposite trends of the wave height Q-Q plots for the two regions gives a hint at the possible cause; the WA are open to mature swell and wind seas that have been generated over long fetch and durations whereas the waves in the NS will be more likely to be younger and generated over shorter fetches and durations. These two wave characteristics relate to two phases in the wave growth physics. It is likely that the wave physics are not imputing energy quick enough in the North Sea during strong wind events. The higher waves in the North Sea will be young wind sea, purely because of the geographical limits of the domain, which means that the wave growth physics have more of a chance to be lagging behind in wave growth compared to a location in the mid-Atlantic where the long fetch lengths means that waves have had considerable time to achieve a fully developed equilibrium with the wind. The issue seems to be exacerbated for waves generated under strong winds (values below the 75%ile are well represented) which suggests that the physics are lacking slightly in their responsiveness to deal with strong and potentially rapidly shifting wind conditions in over short fetches.

The results for the WA show a somewhat opposite trend with the upper quantiles being slightly over-represented, although this over-representation improves again towards the very highest quantiles at the top end of the tail (>10m in the bottom left panel of Figure 3). This appears to be a localised effect as Q-Q results for the overall Atlantic domain exhibit an excellent agreement with the observed climate with only very slight over-representation at higher wave heights (see Figure 13; appendix A).

Deterministic prediction and spread-skill

Figure 5 shows the binned errors in the wave heights (top) and wind speeds (bottom) for the first 6 hours of the forecast. As seen in the previous section, the wind speed errors and biases are very similar between the two regions. Whilst the wave heights RMS errors are fairly similar and follow the same trend, the biases again show an opposite response between the two regions. In the Western Approaches, the binned biases actually follow a similar trend to the wind speed biases with increasing bias at the higher wave heights and wind speeds. However, the North Sea biases are fairly neutral up to approximately 4 metres then become increasingly negatively biased – the opposite trend to the wind biases and WA wave biases.

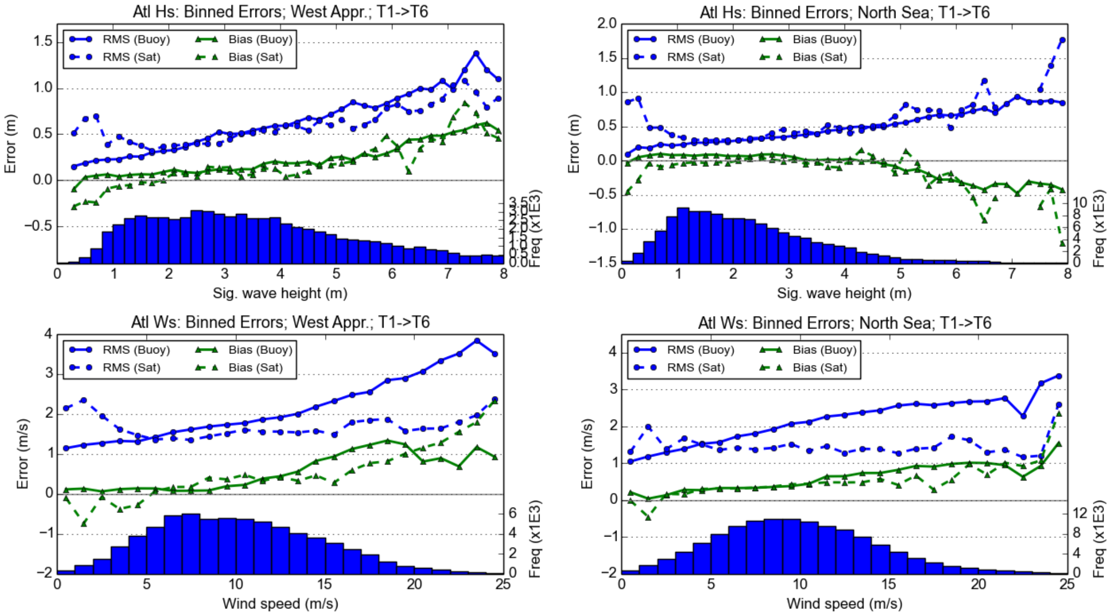


Figure 5 - Bias (green triangles) and RMS (blue bullets) errors for binned significant wave heights (top) and wind speed (bottom) at forecast period T+1 to T+6. Left panels: Western Approaches; right panels: North Sea. Errors derived from buoy (solid lines) and satellite (dashed lines) observations versus model ensemble mean.

It could therefore be argued that the majority of the forecasted wave errors observed in the WA are due to errors in the driving wind model, but errors in the NS are more likely unrelated to wind error and due to limitations in short fetch wave grown in the model wave physics.

Figure 6 shows the overall, 50%ile and 90%ile bias and RMS errors for wave heights wind speed over the entire forecast range. Errors in winds have very similar characteristics in the NS and WA regions, as might be expected from the Q-Q plots presented in the previous section. However, errors in wave heights are again quite different in the two regions. The wave height errors in the NS have a very similar trend to the wind errors (maybe not surprising as this is a region that will have more short fetch generated wind waves). Whilst the overall bias is fairly neutral across the forecast period, the bias of the 50 and 90%iles are increasingly negative which suggests that the model is failing to grow the larger wind waves enough.

The wave height errors in the WA have a very different bias characteristic. The overall, 50%ile and 90%ile biases are all very similar with a slight positive bias between ~0.1 – 0.2m. The RMS errors have a similar trend to the wind speed errors and also the NS

RMS errors, albeit slightly larger. The WA region will be much more open to remotely generated swell and has potentially longer fetch lengths for generation of local waves. This much more mixed wave climate region may be one of the reasons why the bias characteristics are quite different from the wind biases in the Western Approaches.

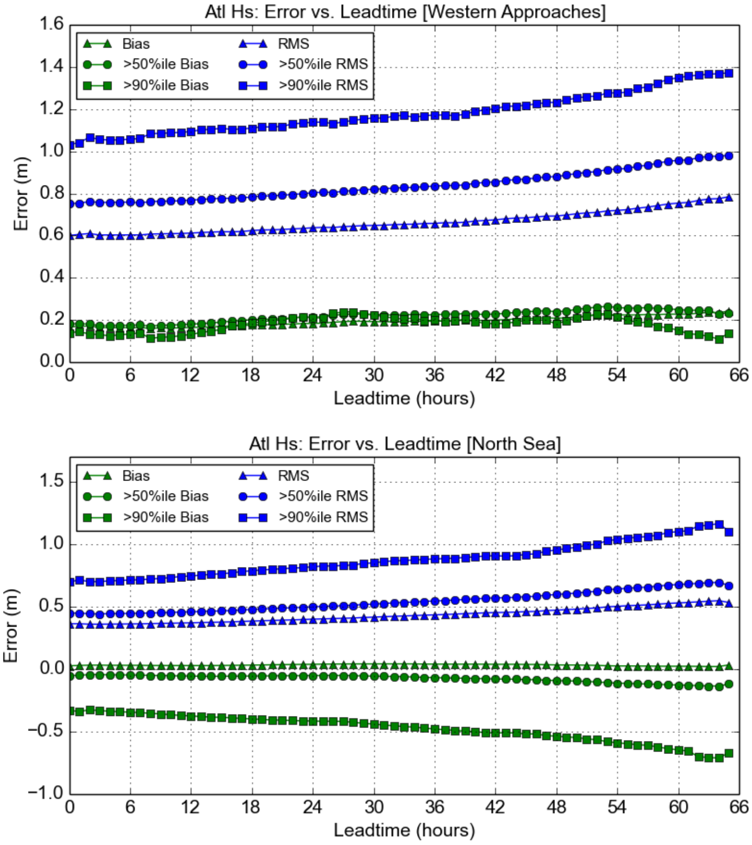


Figure 6 - Significant wave height RMS and bias errors versus leadtime for combined buoy and satellite data. Top: Western approaches; bottom: North Sea. A ±3 hour averaging window has been used.

One of the desirable attributes of an ensemble forecast system is that the spread in the forecast values should ideally encompass the forecast error. This dictates that there should be some relation between the spread and the uncertainty in the forecast. This relationship is most useful at longer leadtimes where there will be more divergence from the initial conditions at the beginning of the forecast and hence more model error. To investigate this relationship, Figure 7 presents the RMS and bias error is plotted against the standard deviation of model spread at day 3 of the forecast (where errors and spread would be expected to be their largest). The histogram in each plot shows the distribution of values in the dataset. It is immediately clear that there is a fairly linear relationship between the spread and both the RMS and bias errors. The wave height spread-skill plots for the WA and NS (top left/right panels respectively) show excellent correspondence between the spread and error with a linear trend for the majority of the range. The tail of the plot is less linear, but it is unclear whether this is due to a change in the error characteristics in this range, or the lower sample population. However, the trend for the NS errors to be slightly higher than the spread and the WA to be slightly lower does suggest that this may be linked to the contrasting error statistics between these regions, as presented earlier.

The fact that the ensemble spread shows a strong correlation with forecast skill suggests that forecasters can use spread to assess model confidence. However, it should be noted that the results of the deterministic verification show that there is also a strong correlation between forecast conditions and skill, i.e. deterministic forecast errors are likely to be higher when storm waves are forecast.

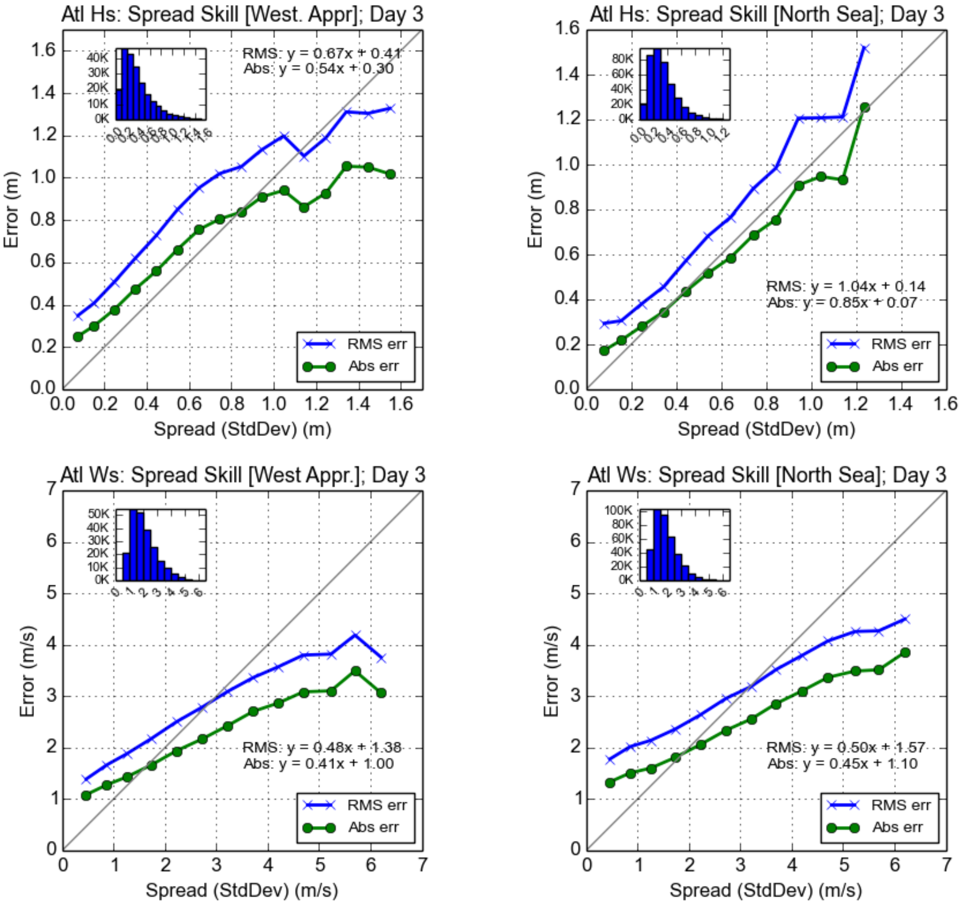


Figure 7 - Spread Skill plots for significant wave height (top panels) and wind speed (bottom panels). Left panels: Western Approaches; right panels: North Sea. Blue crossed line: RMS error; green bulleted line: absolute error. Statistics derived from bias corrected combined buoy and satllite observations.

Probabilistic forecast performance

Ensemble spread

A desirable feature of an ensemble system is for each member in the ensemble to have the same probability of being “correct” with respect to the observations. This means that if each ensemble forecast is ranked according to its value, for a system with good spread the true observation will have an equal chance of falling between any two ranked values. When this process is repeated over many forecasts, a histogram of correct *forecast intervals* is generated. The distribution of this *ranked histogram* (or Talagrand Diagram; see Talagrand and Vautard, 1998) gives a representation of how well the forecast system is reproducing the uncertainty in the observed events. Ideally, a ranked histogram should be flat (i.e. all forecast events occurred with the same frequency). A ‘U-shaped’ distribution means that some observed events do not fall within the spread of the forecast and hence the system is under-spread (however, this assumes no observational error). Conversely, a dome-shaped distribution means that the system is over-spread with the observations only falling between a limited range of members. If the distribution is skewed/asymmetric then the forecast model is most likely biased.

Figure 8 shows ranked histograms for observations in the WA and NS at day 3. The top panels show the results for all observations whilst the bottom panels shoe results for observations exceeding 4 metres only. Although each plot exhibits a ‘U-shaped’ distribution which suggest under-spread, it must be borne in mind that some of the observations may fall outside of the forecast spread due to observational error. To account for this, the black dashed line shows the *ideal model performance* results using the pseudo-observations described previously. Comparing the results to the ideal-performance line reveals that the model is actually well spread with only a small amount of under-spread being evident for all locations and thresholds.

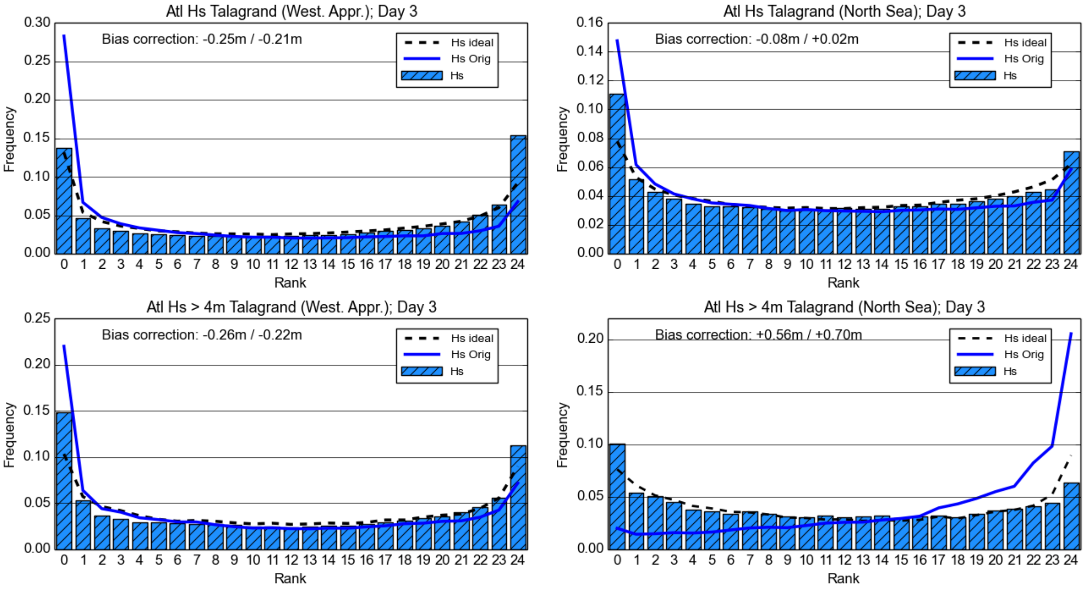


Figure 8 - Talagrand diagrams of significant wave height for the Western Approaches (left panels) and North Sea (right panels). Top panels: all data; bottom panels: data thresholded on 4m observed *Hs*. Combined buoy and satallite observations used (with bias correction). Dashed black line: ideal performance using pseudo-observations; solid blue line: results before bias correction.

Talagrand diagrams are very sensitive to model biases (any systematic model bias will result in the observations being more likely to fall outside either the top of bottom of the distribution) and these biases need to be removed to discern any useful results. The model data used in these Talagrand diagrams has been bias corrected using the overall ensemble mean bias (or in the case of the 4m threshold, the overall ensemble mean bias where the observed H_s > 4m). To show the effect (and importance) of this bias correction, the original results (without bias-correction) are shown as a solid blue line. Here, the opposing characteristics of the 4m exceedance results can be seen between the WA and NS regions with the WA being slightly positively biases (more observations falling below smallest member) and the NS being clearly positively biased (with much of the distribution concentrated at the higher end of the range).

Figure 9 shows similar Talagrand diagrams for the ensemble wind speeds; the bottom panels show the results for a 14 m/s wind speed threshold (being the required wind speed for a fully developed 4 metre wind sea using the Pierson-

Moskowitz (1964) spectral parameterisation). Again, the spread is very good, compared to the results using the pseudo-observations, and is actually slightly better spread than the wave heights. The effects of bias can again be clearly seen in the un-corrected data, but as noted previously, the biases are the same in both the WA and NS regions, unlike the wave forecast.

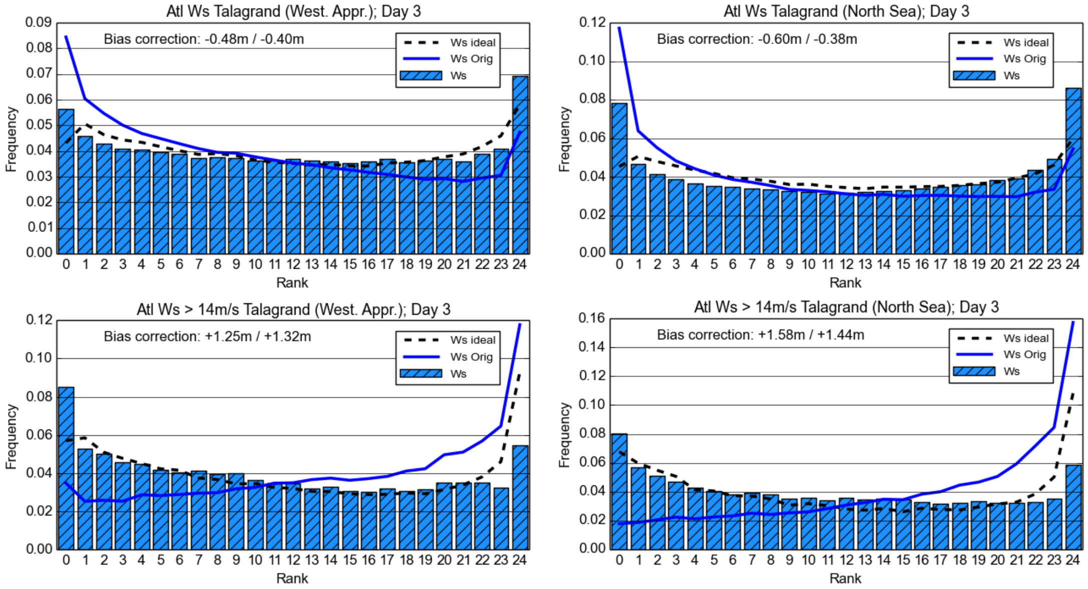


Figure 9 - Talagrand diagrams of 10m wind speed for the Western Approaches (left panels) and North Sea (right panels). Top panels: all data; bottom panels: data thresholded on 14m/s observed wind speed. Combined buoy and satallite observations used (with bias correction). Dashed black line: ideal performance using pseudo-observations; solid blue line: results before bias correction.

Whilst it is important to determine how well spread the model is in relation to the observations, it is also useful to characterise the distribution of ensemble members within the spread of the ensemble itself. Figure 10 shows two plots of the distribution of characteristic ensemble descriptors (mean, median and control) for significant wave height using all sample data (top panel) and only data exceeding 4m H_s (bottom panel).

A 'most likely case' scenario represented by the ensemble median is most likely to fall close to, but slightly below, the ensemble central H_s value in terms of data range between predicted maximum and minimum. Observed outcomes are more widely distributed through this range and have a strong chance of falling above the central value, or outside the predicted range altogether. Note that the mean, median and control distributions are much more central in the Hs > 4 distribution (bottom panel of Figure 10). This is likely due to the characteristics of the waves that exist above 4m; they will be almost entirely wind driven, whereas the distribution for all sample data will include smaller, but remotely generated swell waves.

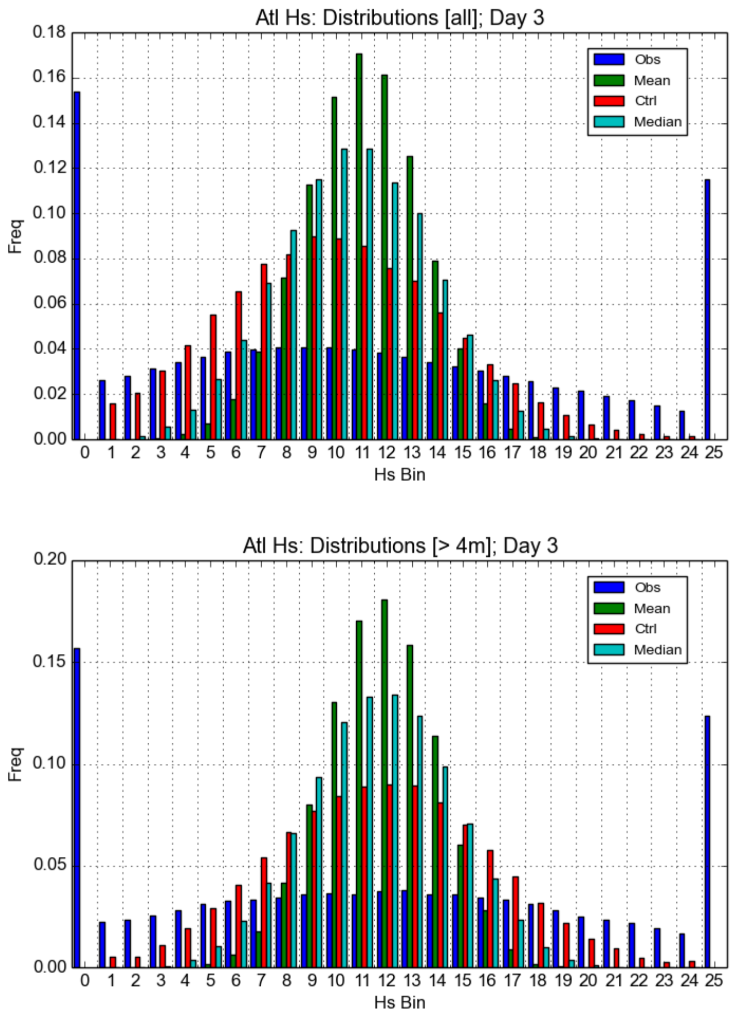


Figure 10 - Occurrence of observations (blue) and selected 'members' (mean-green, control-red, median-blue) from ensemble within (bias corrected) ensemble prediction H_s range. The bin values 1-24 represent equally spaced H_s ranges between ensemble prediction minimum and maximum; bins 0 and 25 contain values falling outside the ensemble range. (Top panel) Using all data in verification sample, (bottom panel) data sample based on observation falling above 4m.

Overall, the wave ensemble is on average slightly under spread – a characteristic that it has in common with MOGREPS-G. The spread in the wave system is entirely inherited from the driving atmospheric model, therefore any under-spread in the winds can be expected to impact the spread in the downstream wave model. In forecasting terms the impact will be that the probability of observations exceeding the 'reasonable worst case' scenario defined by the ensemble rank maximum is higher than might be expected.

Threshold exceedance reliability

As ensemble forecast give a probability of an event occurring, rather than a binary "yes/no" prediction, a different approach to verification of exceedance of thresholds needs to be undertaken. One commonly used metric is the *reliability diagram* which compares the forecasted probability with the observed frequency for a set of probability categories. For example, for forecasts where the model predicts that a threshold is exceeded with a probability of 20% it would be expected that 1 in 5 of the observations should also show an exceedance. For this report, 24 evenly spaced probability categories have been used (as there at 24 members in the total lagged ensemble). A *reliable* forecast should ideally have points lying along the diagonal line on the reliability diagram (see Appendix B). However, as shown in the Talagrand diagrams, the existence of observational error will always preclude this so an *ideal model-performance* line is also plotted using the same pseudo-observations as previously described.

As systematic model biases will affect the threshold exceedance predictions, the model data has been bias corrected prior to verification. A slightly different bias correction approach has been adopted in the case compared to the Talagrand bias correction. As the exceedance reliability metric is based on the exceedance of single value, a *local* bias correction has been applied where binned bias error around the threshold value has been used rather than the overall bias error of the entire wave height range. Binned wave height and wind speed errors for day 3 are presented in Figure 14 (Appendix A).

Figure 11 shows the 4 and 6 metre wave height exceedance reliability for the WA and NS. The population in each probability category is shown in the inset histogram. After bias correction, it is clear that the forecasts exhibit a high degree of reliability for both of the exceedance thresholds. There is some evidence of slight under-spread in the WA (which manifests as slight under-confidence at low probabilities and over-confidence at higher probabilities, resulting in an S-shaped curve), but this is only slight when compared to the ideal model performance (dashed gray line).

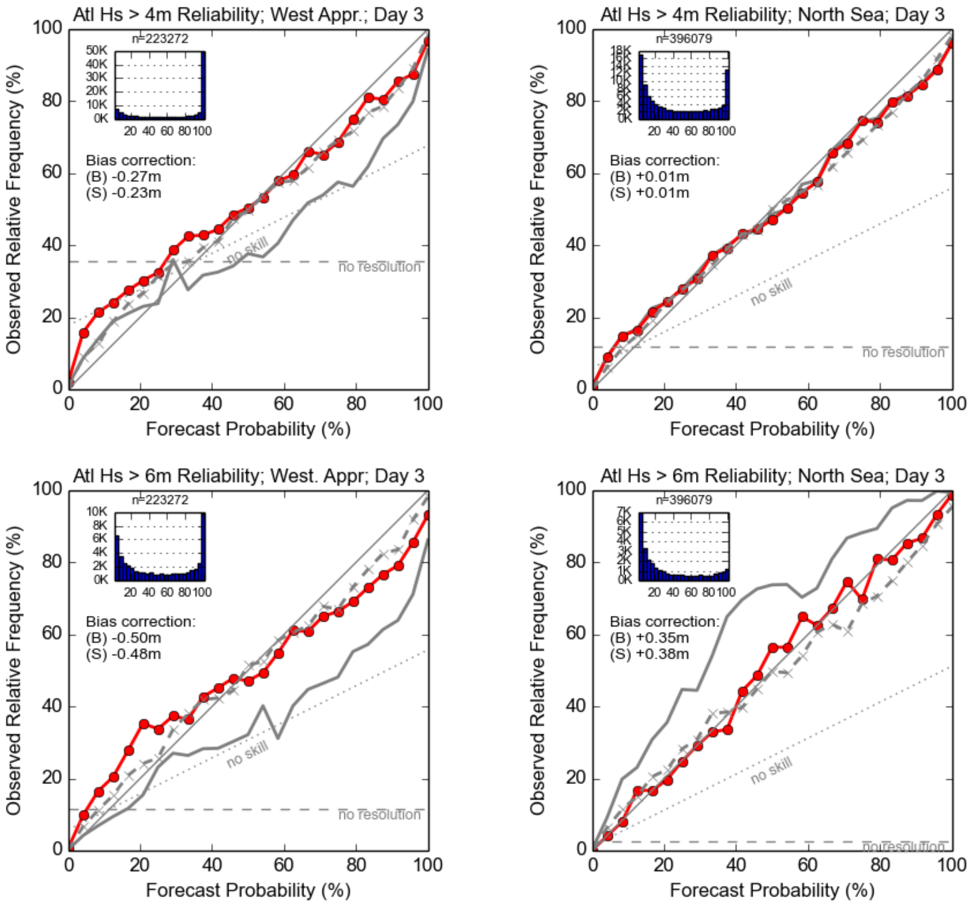


Figure 11 - Significant wave height exceedance reliability plots (red lines+bullets) for *H_s>4m* (top) and *H_s>6m* (bottom). Left panels: Western Approches; right panels: North Sea. Reliability derived from bias corrected combined buoy and satellite data. Solid gray line: reliabiltiy of uncorrected forecast data. Dashed gray line: verification against peudo observations.

Figure 12 shows the reliability plots for wind speed exceedances of 14m/s and 17m/s. These are the wind speeds required to maintain a fully developed Pierson-Moskowitz spectrum with a H_s of 4m and 6m, respectively. Again, the forecasts exhibit good reliability with some evidence of slight over-confidence at the higher probability categories. It can again be noted how the uncorrected wind speeds are over-confident in the North Sea, yet the uncorrected wave heights have a tendency to be under-confident.

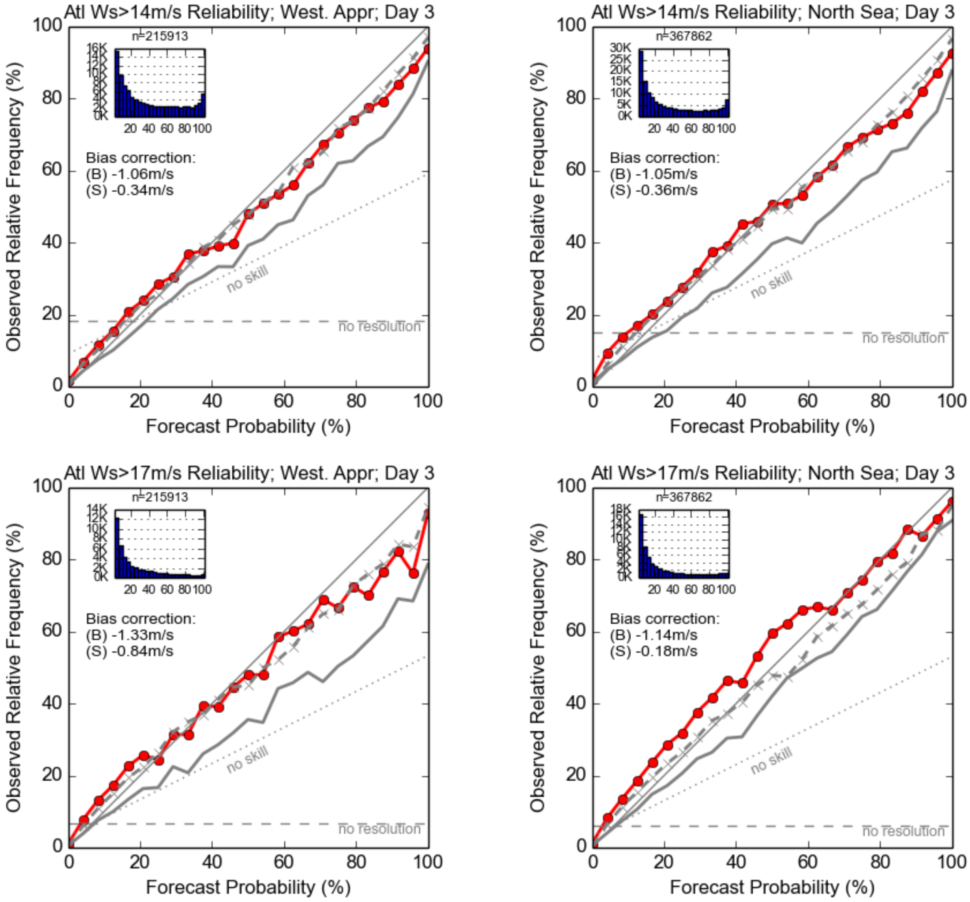


Figure 12 – 10m wind speed exceedance reliability plots (red lines+bullets) for *Ws>4m/s* (top) and *Ws>6m/s* (bottom). Left panels: Western Approches; right panels: North Sea. Reliability derived from bias corrected combined buoy and satellite data. Solid gray line: reliabiltiy of uncorrected forecast data. Dashed gray line: verification against peudo observations.

Conclusions

Verification statistics for the 24 member Atlantic wave ensemble system have been presented for the Western Approaches and North Sea areas of the U.K. These two areas have been verified separately as they have different wave characteristics; the Western Approaches are open to the Atlantic and receive mature waves generated over potentially long fetches and durations whilst the North Sea is sheltered from the Atlantic and will be more likely to be dominated by fetch limited wave growth conditions.

Whilst overall statistics for the whole Atlantic domain show virtually nil bias, there are small regional biases present in the two U.K. areas detailed in this report. As the Talagrand and reliability metrics are sensitive to systematic biases, a bias correction has been applied to the model data before verification using these metrics. Some of the wave height biases can be attributed to similar biases characteristics in the wind, but there is a clear contrast in the wave biases between the two regions with the Western Approaches tending to be biases slightly high whilst the North Sea being biased low. This is due to the wave physics not imputing energy quick enough in the fetch limited North Sea area (this is also likely to be an issue in the Irish Sea); this results in a lag in the wave growth which shows clearly in the negative bias of the larger wind sea waves ($H_s > 4$m). This bias characteristic is not seen in the Western Approaches as this area is open to the Atlantic where the waves have had sufficient time and fetch to develop to an equilibrium state with the driving wind.

Both areas around the U.K. show good spread in both the wave heights and wind speeds. There is also a clear relationship between the ensemble spread and the model error which means that the model spread is a good indicator of forecast skill and is likely to be strongly correlated with the local conditions. There is some evidence of under-spreading in the Talagrand plots (U-shaped distribution), but the use of pseudo-observations derived from the ensemble has shown that a significant amount of this under-spread is due to the presence of observational errors.

The use of reliability plots has shown that the system has excellent skill (once a simple systematic bias correction is applied) at predicting probability of threshold exceedance events in both the wave heights and wind speed.

The choice of Tolman & Chalikov (1996) physics for the Atlantic model was based up the performance of open water wave generation and effective swell dissipation in the tropics. However, it has been identified that these physics are potentially not dynamic enough to grow waves quick enough in short fetch areas or rapidly changing wind conditions. It is therefore planned to transition to the WAM4 physics in early 2015. Tuning tests of the WAM4 physics are currently being undertaken to ensure acceptable performance of that the swell dissipation in the tropics.

Recommendations

Based on the results of this study and the wider MyWave work the following recommendations are made for implementation of the final wave ensemble system:

1. The ensemble systems show sufficient skill for Met Office to support continued running and product development – subject to PWS resource for the science team and business area resource to IT systems teams in production.

2. The Atlantic model is given priority status in terms of implementation and development – the model should be extended to 7 days for consistency with ensemble surge, and should be updated to use either the TS3M or TS3H flavour of WAM Cycle-4 physics in order to resolve North Sea and Irish Sea bias issues. We plan to implement these updates in PS35 (Nov-Jan 2014-15).
3. The results of the verification are relevant for ‘offshore’ wave conditions. It should be noted that inshore water levels are a strong control on wave variability and further exploration on the effect on the wave ensemble spread in the nearshore is undertaken within the demonstration project.
4. Whilst recognising EA constraints and forecasting procedures, the results from the wave verification supports the Met Office position (established for other ensemble models) that use of probabilities over multiple thresholds is the ‘best practise’ method for decision making from the ensemble. The use of bias correction when creating probability products is recommended.

References

Bidlot, J.-R., 2012: Present status of wave forecasting at E.C.W.M.F. In Proc. ECMWF Workshop on Ocean Waves, Reading, 2012, p1-15.

Bunney, C. C., Li, J-G., & Saulter, A. (2012) Met Office Wave Model Ensemble Prediction Systems in the 'Atlantic-Euro Zone'. *MyWave Report D3.1*. December 2012. Met Office, U.K.

Bunney, C. C. (2014) Performance and verification of the Mt Office “Atlantic-Euro Zone” ensemble wave model. *MyWave Report D3.5*. June 2014. Met Office, U.K.

Li, J. G. 2012: Propagation of Ocean Surface Waves on a Spherical Multiple-Cell Grid. *J. Comput. Phys.*, **231**, p 8262 - 8277.

Palmer, T. & Saulter, A. (2013) Estimation of regional observation errors and application to MyWave metrics. *MyWave report D4.3*. December 2013. Met Office, U.K.

Pierson, W. J. & Moskowitz, L. (1964) A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigorodskii”. *J. Geophysical Res.* **69,** 5181 – 5203

Saetra, Ø., Hersbach, H., Bidlot, J.-R. & Richardson, D. (2004b) Effects of Observational Errors on the Statistics for Ensemble Spread and Reliability. *Mon. Wea. Rev.,* **132**, 1487–1501.

Swets, J. A., (1973) The relative operating characteristic in psychology. *Science*, **182**, 990-999.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proceedings: ECMWF Workshop on Predictability*, ECMWF, 1–25

Tolman, H.L. and D. Chalikov, 1996: Source terms in a third-generation wind-wave model. J. Phys. Oceanogr., 26, 2497-2518.

Wilkes, D. S. (2001) A skill score based on economic value for probability forecasts. *Meteorol. Appl.* **8**, 209 – 219.

Appendix A – Extra plots

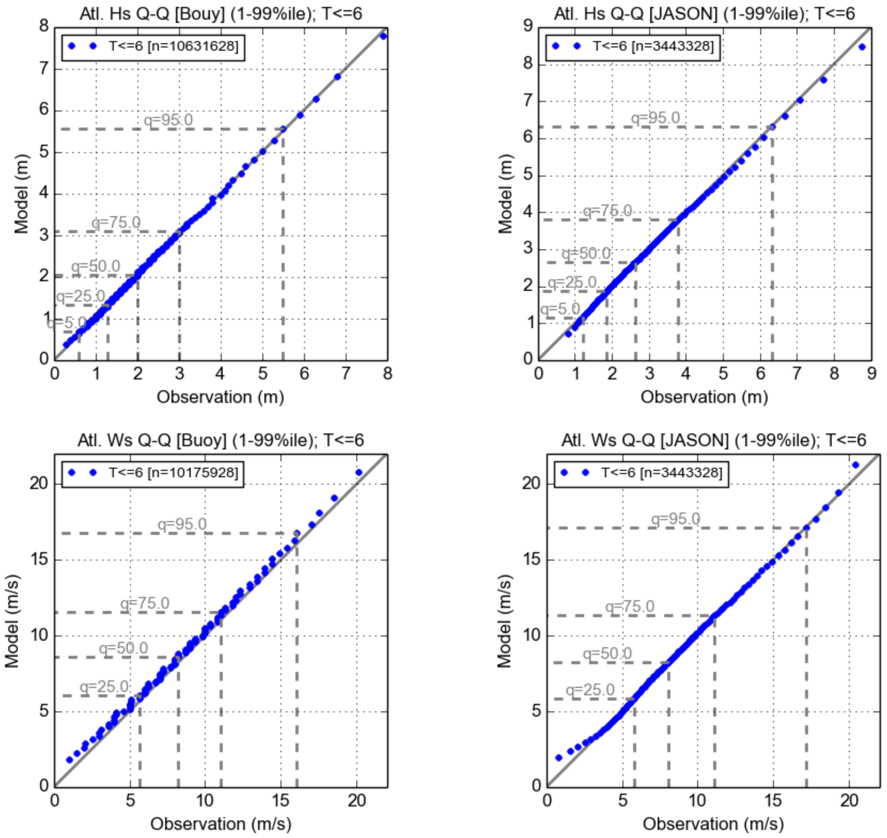


Figure 13 - Quantile-Quantile plots of Atlantic model versus observation for 1 – 99 percentiles. Top row: significant wave height; bottom row: wind speed. Buoy observations are shown on the left, JASON2 observations on the right. Dashed lines show the 25th, 50th, 75th and 95th percentiles.

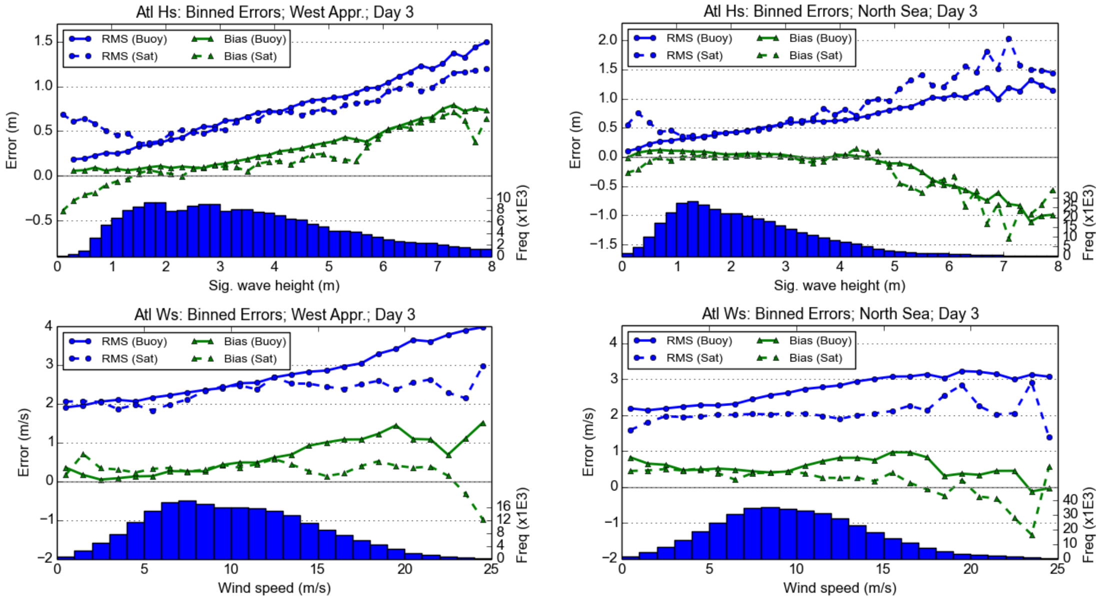


Figure 14 - Bias (green triangles) and RMS (blue bullets) errors for binned significant wave heights (top) and wind speed (bottom) at forecast period T+1 to T+6. Left panels: Western Approaches; right panels: North Sea. Errors derived from buoy (solid lines) and satellite (dashed lines) observations versus model ensemble mean.

Appendix B – Description of metrics

Reliability diagrams

Reliability Diagrams compare the forecasted probability of an event with the observed frequency. The reliability curve is constructed using multiple forecast probability thresholds (in the wave ensemble, thresholds are spaced evenly by the number of ensemble members, i.e. probability increments of *1/24 = 0.042*). Perfect model reliability results in a straight *y = x* line along the diagonal. The effects of model error and under/over-spreading usually result in some deviation from this ideal; examples of typical reliability plots are shown in Figure 15 [from Wilkes, 2005]. The inset histogram shows the population of each probability bin and is a useful for determining if a poor reliability at a particular threshold is linked to under-sampling (common in high probability of rare event thresholds).

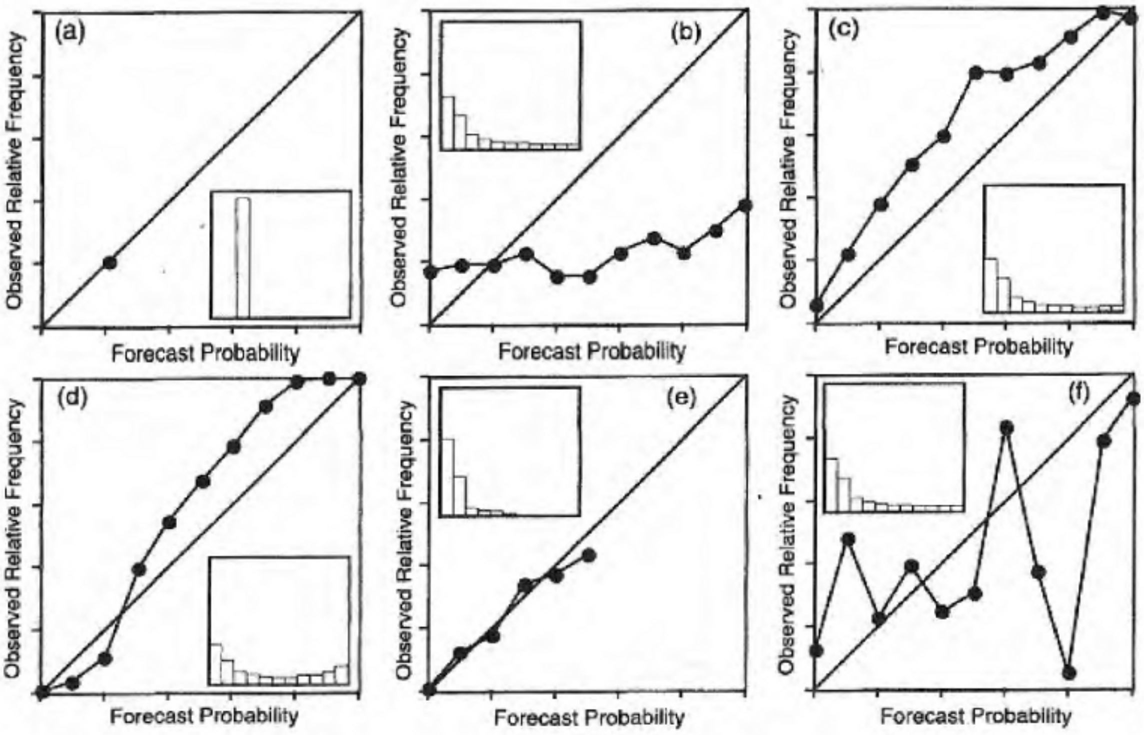


Figure 15 - Hypothetical examples of reliability diagrams: a) climatological forecast, b) minimal resolution, c) under-forcasting bias, d) good resolution at expense of reliability, e) reliable forecast of rare event, f) sample size too small. Inset boxes show population of forecast probability bins. [From Wilkes, 2005]

Met Office
FitzRoy Road, Exeter
Devon EX1 3PB
United Kingdom

Tel (UK): 0870 900 0100 (Int) : +44 1392 885680
Fax (UK): 0870 900 5050 (Int) :+44 1392 885681
enquiries@metoffice.gov.uk
www.metoffice.gov.uk