MyWave

Estimation of regional observation errors and application to MyWave metrics

Reference: MyWave-D4.3

Project N°: FP7-SPACE-2011-284455	**Work programme topic:** SPA.2011.1.5.03 – R&D to enhance future GMES applications in the Marine and Atmosphere areas
Start Date of project : 01.01-2012	**Duration**: 36 Months

WP leader: Andy Saulter	**Issue:** 1.0
Contributors : Tamzin Palmer, Andy Saulter	
MyWave version scope : All	
Approval Date : 23 Dec 2013	**Approver:** Andy Saulter
Dissemination level: Project	

DOCUMENT

VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:	Andy Saulter	WP4	23 Dec 2013
Distribution			
	Ø. Saetra (Project coordinator) A. Saulter (WP4) J.-R. Bidlot (WP4) M. Gomez-Lahoz (WP4) T. Palmer (WP4)		

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By
0.1	17 Dec 13	all	First draft of document	Tamzin Palmer	Andy Saulter
1.0	23 Dec 13	all	Document finalization	Tamzin Palmer	Andy Saulter

TABLE OF CONTENTS

LIST OF FIGURES

Figure 3.14. Observation prediction skill comparisons for Success Ratio against a 2m threshold at varying model lead time and idealised prediction. X-axis data are comparisons against in-situ observation and y-axis data are comparisons against Jason-2 altimetry.

LIST OF TABLES

GLOSSARY AND ABREVIATIONS

Baseline prediction	Prediction system used as a source of verification comparison
ECMWF	European Centre for Medium range Weather Forecasts
Hs	Significant wave height
IOC	International Oceanographic Commission
JCOMM	Joint Commission on Marine Meteorology
MAE	Mean Absolute Error
MCS	Marine Core Service
NEAM	North European Atlantic Margin
pdf	Probability distribution function
Q-Q	Quantile-quantile (plot)
Reference data	Data used to verify a prediction
RMS	Root Mean Squared (value of parameter)
(R)MSE	(Root) Mean Squared Error
SI	Scatter Index
SNRMSE	Symmetrically Normalised Root Mean Squared Error
WMO	World Meteorological Organisation

APPLICABLE AND REFERENCE DOCUMENTS

Applicable Documents

	Ref	Title	Date / Issue
DA 1	MyWace-A1	MyWave: Annex I – "Description of Work"	September 2011

Reference Documents

	Ref	Title	Date / Issue
DR 1	MyWave-D4.1	MyWave: Definition of experiement plan and resources for MyWave Task 4.1: Identify 'compatible metrics' using remote sensed and in-situ wave measurement baselines	September 2012 / v1.0
DR 2	MyWave-D4.2a	MyWave: Proposal of metrics for user focused verification of deterministic wave prediction systems	October 2013 / v1.0

I INTRODUCTION

MyWave WP4 seeks to define operational verification methods that can be robustly applied within a wave element of a Marine Core Service (MCS). In particular, project tasks have been concerned with methods associated with the provision of consistent information regarding wave forecast model uncertainty based on verification that uses a mix of both satellite remote sensed and in-situ observations of the true sea-state as a reference. The aim is to describe sampling properties, representation scales and observation errors associated with the two types of observing system, and to assess and quantify variability in metrics derived when wave model outputs are verified using these baselines. The outcome from WP4 is to propose a set of measures that will ensure a MCS for waves can provide a set of self-consistent performance metrics across European waters for primary marine forecast parameters such as significant wave height and mean wind speed using either, or both, in-situ and satellite observations as a reference.

In this report a regional assessment of measurement errors associated with in-situ platforms and satellite altimetry has been made (Subtask 4.1.1), and a method to apply the resulting data within verification metrics is demonstrated and discussed (Subtask 4.2.1). The focus is on establishing a generic and user relevant approach to verification which can be reviewed through generation of example metrics and ongoing consultation with end users (Subtask 4.2.1) in order to produce a final proposal for essential components of a MCS wave verification system (MyWave D4.4). Section 2 presents the results of the regional triple collocation study. Section 3 discusses the verification methodology and Section 4 provides a summary and notes follow-up activities for WP4 during the remainder of the MyWave project.

II ESTIMATION OF REGIONAL OBSERVATION ERRORS USING A TRIPLE COLLOCATION METHOD

In order to account for observation errors within verification, those errors must first be quantified. Whilst observation error statistics have been derived for global applications using triple collocation methods (e.g. Janssen et al., 2007), a sufficient body of evidence exists to suggest that regional variations in these statistics are likely to occur, particularly in respect to the errors attributed to in-situ networks (e.g. Durrant et al., 2009). In this study the feasibility of quantifying observation errors specific to regional sea areas has been explored.

II.1 Study regions and source data

Regional triple collocation assessments of observation errors in measurement of significant wave height were carried out for two European sea areas where high densities of in-situ data are available:

- North Sea (3°W - 10°E , 51°N - 63°N)
- North European Atlantic Margin (NEAM, 20°W- 0°W, 30°N- 65°N)

These regions represent somewhat different environments in terms of the wave climate, with the North Sea area being a semi-enclosed shelf sea generally dominated by short to moderate fetch wind-seas, whilst waves in the NEAM have often developed over longer fetches and the wave climate comprises a mix of developing and mature wind-seas plus swell.

In order to use contemporary data from three independent sources, model, buoy and satellite the period from 2010 to 2012 inclusive were selected for this study. The wave model data used came from a hindcast run using the WAVEWATCH III model (Tolman, 2009) configured for an 8km resolution European domain. In-situ data were sourced from an hourly observation dataset made available to ECMWF as part of the WMO/IOC Joint Commission On Marine Meteorology (JCOMM) international wave forecast intercomparison project (Bidlot et al., 2007). Fast delivery satellite altimeter data from Envisat, Jason-1 and Jason-2 missions were downloaded for this period via the GlobWave project (Globwave, 2012).

II.2 Study Method

II.2.1 Error estimation method

A number of examples of global triple collocation studies are available from the literature and have been described in report MyWave-D4.1 (section 3.1.1). The error assessment method used in this study follows that of Janssen et al. (2007; hereon denoted as JEA07, see also MyWave-D4.1 Section 3.3.2), although details of the spatial and temporal windows used for collocation have been changed in order to reflect regional correlation lengthscales and the higher resolution of the regional wave model used (see next subsection). The method assumes that observations of true sea-state comprise a systematic error and an (independent) random error. Outputs from the analysis estimate these error components using (respectively) a linear calibration constant (slope) and relative error (Scatter Index, SI) value. The slope is calculated relative to one data source as an unbiased estimator of the truth and in this study, consistent with JEA07, the in-situ data were used as that reference observation.

II.2.2 Collocation criteria

Choosing suitable spatial and temporal scales for collocation of the model and measurements is a crucial part of any triple collocation study. The scales chosen will depend on the spatial resolution of the model, the spatial distribution of the in-situ data and the numbers of collocations available. While it is desirable to collocate the measurements as closely as possible in space and time, it is also essential for the error estimates to be derived from a statistically robust dataset. Therefore a balance needs to be struck between using the largest possible sample and ensuring that the collocations are valid.

In order to investigate the spatial scale at which observations made at different locations will measure conditions that are sufficiently similar to allow an assumption that the observing errors are directly comparable, a study using the background error covariance matrix was carried out in the two study areas. The details of this study are given in Palmer and Saulter (2013). The results of this investigation were used to establish spatial distances at which high levels of correlation in background error (greater than 0.8) occurred relative to different in-situ platform locations. The lengthscales varied with location, but were generally longer in the NEAM and shorter in the North Sea. The averaged results are shown in Tables 2.1-2.2 and generally support using a collocation distance of 50km. It should also be noted that this study related to the correlation of background errors and, with further investigation, actual

Table 2.1. NEAM 0.8 correlation distance (km) by season and directional sector.

	Spring	Summer	Autumn	Winter
North	104.8	70.2	100.9	88.3
South	112.7	71.6	115.2	84.8
East	152.3	106	149.5	116.7
West	156.4	108.3	159.8	115.0

Table 2.2. North Sea 0.8 correlation distance (km) by season and directional sector.

	Spring	Summer	Autumn	Winter
North	49.2	36.4	45.0	60.7
South	44.3	35.0	43.7	52.8
East	85.0	75.0	91.4	94.0
West	81.4	72.9	92.8	115.0

Table 2.3. NEAM 0.5 correlation distance (km) by season and directional sector.

	Spring	Summer	Autumn	Winter
North	230.7	160.2	234.1	208.9
South	281.6	180.0	286.3	244.8
East	336.7	274.5	356.4	327.6
West	380.5	256.0	408.7	353.7

Table 2.4. North Sea 0.5 correlation distance (km) by season and directional sector.

	Spring	Summer	Autumn	Winter
North	145	107.1	116.4	180.0
South	120.7	85.7	250.7	177.1
East	194.3	117.5	221.4	222.0
West	181.4	172.9	211.4	279.2

measurements were often found to be highly correlated over much greater distances. Some notable exceptions were found in the North Sea, but this was largely due to the proximity of in-situ platforms to the coast, where gradients in the spatial structure of the wave field are high and satellite measurements can be unreliable. These locations were removed from the triple collocation study.

A 0.5 correlation distance was also assessed in order to provide an indication of the distances at which individual measurements might be considered to be independent. Establishing an independence lengthscale was considered an important test for the study since, if the same conditions are sampled disproportionally, then these data may skew the results of the triple collocation error estimate. Averaged distances for the 0.5 correlation of background errors are shown in Tables 2.3 and 2.4. These results show that in some areas the proximity of individual wave buoys may result in a number of duplicate measurements existing within the triple collocation sample, particularly in the northern North Sea. In order to test the sensitivity of the results to the location of these wave buoys some subsampling of the data was carried out.

II.2.3 Sensitivity tests

Due to the constrained areas and time periods necessitated by undertaking regional error assessment of contemporary observations, a key difference between this study and JEA07 is the number of data available in the triple collocation sample. In order to make some evaluation of the robustness of the resulting error assessments a number of sensitivity tests were made.

One test was to establish the effect on the results of the number of satellite soundings averaged in order to provide the matchup satellite value (or so-called 'super-observation'). Error assessments were carried out using super-observed data comprising 3 and 5 altimeter soundings within a given super-observation and contrasted with a control dataset that used a single sounding.

The stability of the errors calculated using the triple collocation method was assessed using a rolling analysis of 12 month data samples over the available satellite data within 2010-2012. The start time of the data sample was moved forward by one month at a time. A sampling window of one year was used to ensure that a sufficiently large number of matchups occurred in each sample. In the NEAM typical sample sizes obtained were of the order of 550-750 data values for the Jason-2 satellite and 1000-2000 values for all satellite

data combined during the period 2010-2012. In the North Sea Jason-2 sample sizes were in the region 750-1200 and 1200-1600 data values were captured when all satellite observations were combined. These samples sizes enabled convergence of the error estimation algorithm, consistent with findings of other researchers (Peter Janssen, *pers. comm.*).

The requirement for the matchup data to be independent was tested using data subsamples based on the independence correlation lengthscale established in Palmer and Saulter (2013). The distance allowed between each matchup (based on the in-situ platform location) was expanded from 50km to 100kmand 150km and in the NEAM also to 200km, 250km, 300km and 350km. A time window of 3 hours was used so that if the samples occurred during a very different time period (e.g. on a different day) they would not be excluded from the data set. It was found that in the NEAM this had little impact on the results because matchups did not occur within 300km and 3 hours of each other. This is not unexpected due to the size of the domain and the distance between the individual in-situ locations. In the North Sea however there are a number of in-situ platforms located in relatively close proximity and the subsampling resulted in a modest reduction of the number of matchups used in the analysis. Results from analyses carried out on the subsampled dataset were compared with results generated when the full sample of matchups was used.

II.3 Study Results

II.3.1 Sensitivities to inhomogeneity in the matchup sample

In the assessment of 12 month rolling data samples the Jason-2 satellite data were treated as a special case since this mission provided a robust sample of data throughout the study period and thus provided a consistent dataset against which to examine the temporal stability of the error estimates. Figure 2.1 shows an example time-series of error estimation results (SI and slope) for this satellite in the NEAM. In this analysis the satellite data were super-observed such that three altimeter soundings were averaged in order to generate the satellite data value, and each month given on the x-axis in the figure is the first month of a 12 month sample (i.e. January 2010 corresponds to a sample taken over the period January-December 2010). In Figure 2.1 it is immediately evident that there is a significant change with time for the in-situ and model errors and some more limited instability in the Jason-2 data also. The results are more stable in the latter part of the series (post January 2011) however. This

raised the question as to whether these changes were brought about by in-homogeneities in the data sample, or as an effect of sample size itself. In reviewing the model data it was noted that, for the hindcast used, a change in the wind data source was made in January 2011 (from a downscaling atmospheric model run using ERA-Interim boundary conditions, to operational Met Office global atmospheric model analysis winds) which was likely to have improved the correlation of the wave model with observations. This is consistent with the steady fall in the model error during 2010, since with each successive 12 month window more post-January 2011 data will be incorporated in the sample, and subsequent levelling out of the model SI and slope values.

The sharp change in SI for the in-situ data was also linked to an anomaly in the data sample. Figure 2.2 shows Hs scatter plots for in-situ versus Jason-2 data and the locations of in-situ sites within the sample for each of the 12 month windows used in the analysis. Common to these plots, prior to the December 2010 sample, are the existence of some outlying high in-situ Hs readings matched with significantly lower satellite (and model) values and the presence of the Brittany wave buoy. When this location is no longer present in the data record (post December 2010) the in-situ errors reduce and stabilise. To test that these few outliers were responsible for the changes in the in-situ error estimates, the data were removed (4 data values were identified at Brittany) and the analysis re-run. Figure 2.3 shows the revised results. In-situ errors were stabilised for the whole period and reduced in 2010-2011 data by approximately 4%. Jason-2 slope and SI were also stabilised (variability of order 1%), whilst the model errors retained their steady improvement to a stable state during the analysis period. A similar effect was noted when enlarging the dataset by including all available satellite data (sample size increased by a factor of 2-3) and suggests that although convergent error estimates were achieved, the triple collocation method can be particularly sensitive to a few outliers in the sample and that quality control procedures in the assessment need careful attention. On a more positive note, this example has also shown that the technique is extremely adept at identifying issues in a small part of the observing network and that with proper quality control the change in model skill could be tracked.

A similar analysis was conducted for the North Sea data and the results are shown in Figure 2.4. In this instance the satellite error estimates are relatively stable, whilst a steady drift (toward lower errors) are seen in the rolling 12 month samples of both model and in-situ data. For the model the drift is consistent with the result found the NEAM and the change in hindcast wind forcing. The drift in the in-situ errors appears consistent with changes in the in-situ network, but may also be influenced by the presence of a few particularly large storm

events within the match-up sample as of January 2011 (Figure 2.5). These effects can be seen as a particularly marked change in the slope value for the Jason-2 data, which can be attributed purely to in-situ data changes as it is not believed that the altimeter observation processing algorithm for the satellite was altered dramatically during this period.

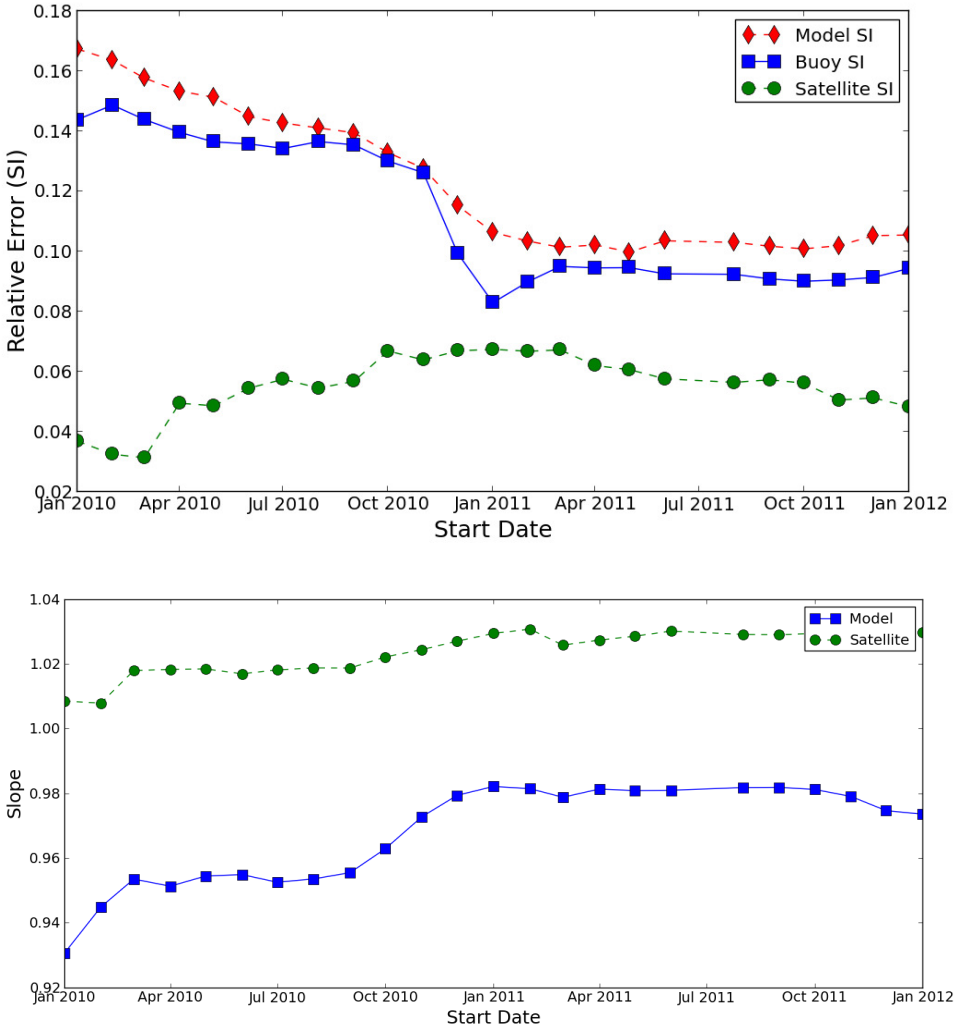


Figure 2.1. Time-series of rolling 12 month sample error estimates for model, in-situ platform (labelled buoy) and satellite SI (upper panel) and slope (lower panel) for the NEAM. Samples were determined based on Jason-2 altimeter data with 3 soundings in each super-observation; the in-situ data provided the reference 'truth' for the slope estimate.

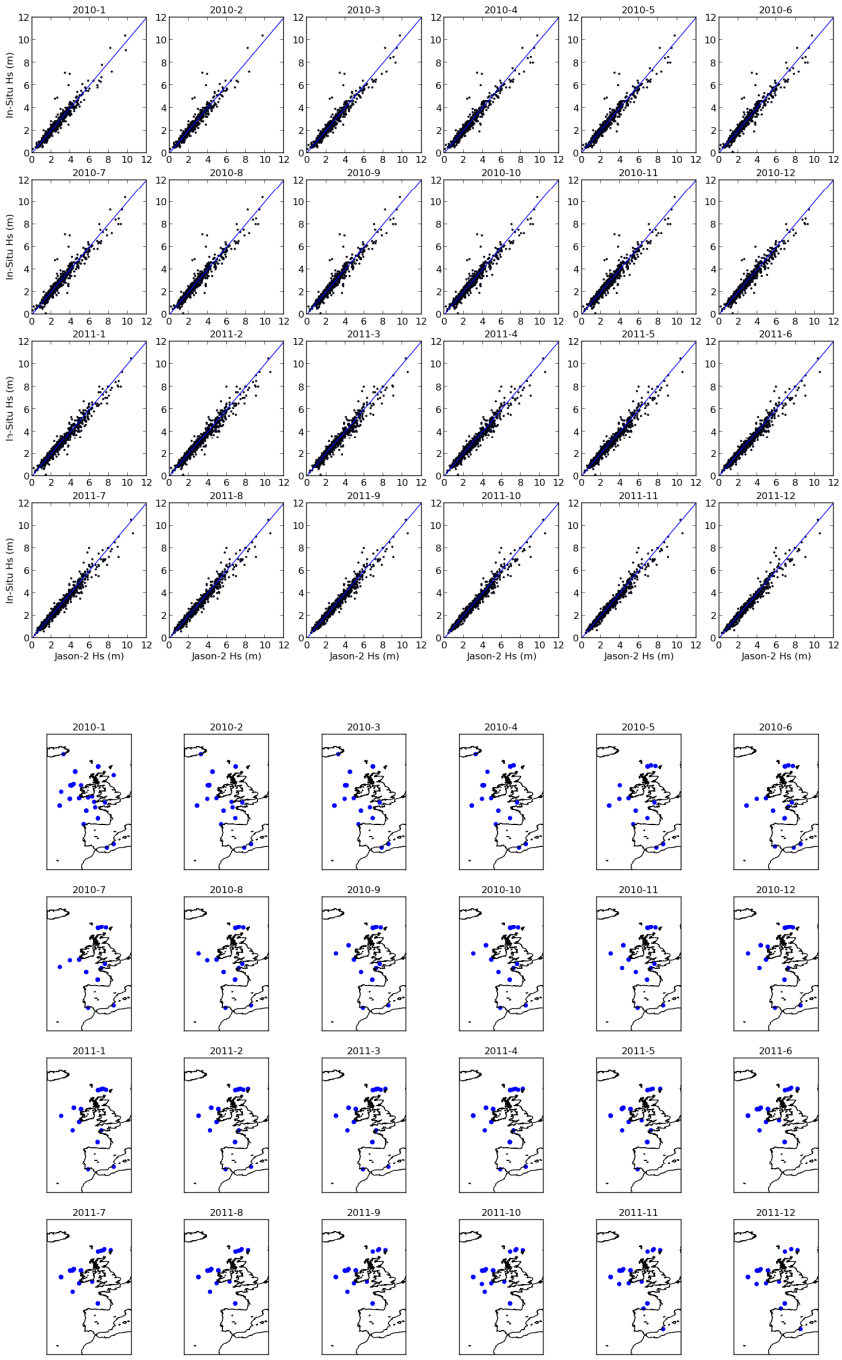


Figure 2.2. Hs scatter data (Jason-2 vs in-situ, top panel) and unique in-situ locations used in rolling 12 month triple collocation samples for NEAM analyses.

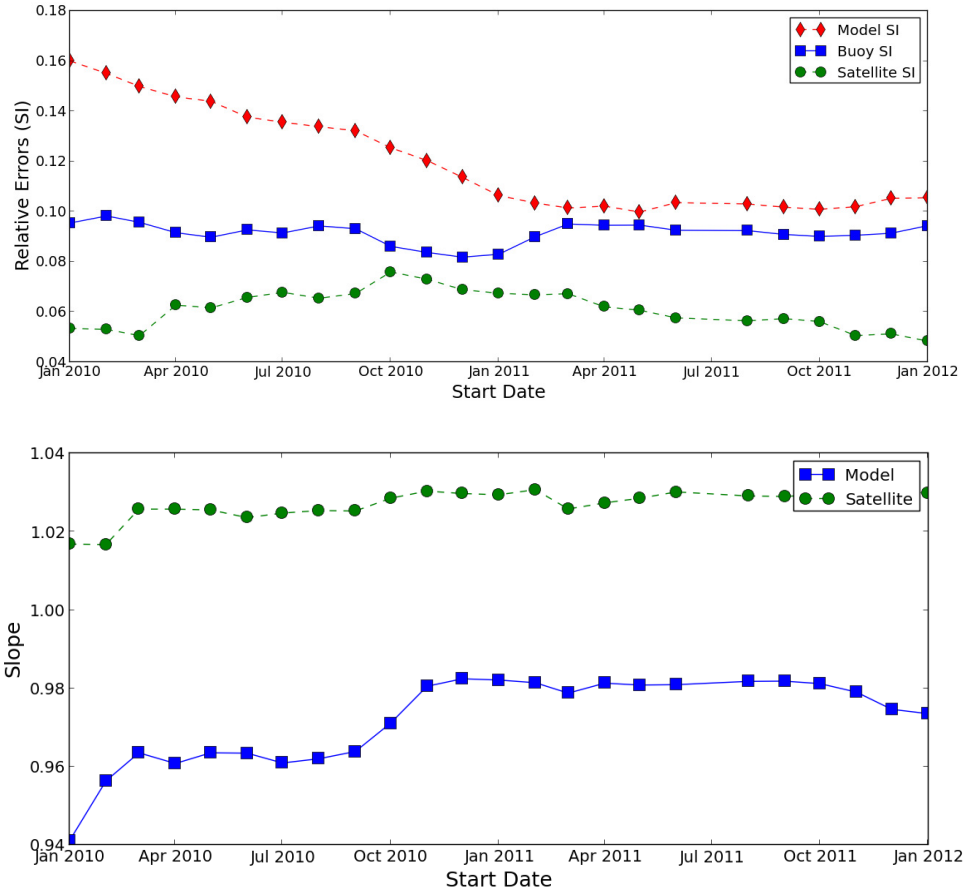


Figure 2.3. Time-series of rolling 12 month sample error estimates for model, in-situ platform (labelled buoy) and satellite SI (upper panel) and slope (lower panel) for the NEAM after removal of Brittany outlier data. Samples were determined based on Jason-2 altimeter data with 3 soundings in each super-observation; the in-situ data provided the reference 'truth' for the slope estimate.

Figures 2.6 and 2.7 illustrate the sensitivity of the error estimates to varying the satellite data contributing to the matchup sample by presenting time-series of SI and slope for an analysis in which all available satellite observations were used. Figure 2.6 shows the NEAM analysis and can be compared directly with Figure 2.3. The effect of including the extra match-ups is to increase SI estimates for all three data sources by approximately 1%, whilst the slope values for both model and satellite are increased by approximately 2%. For the North Sea data in Figure 2.7, the change from the Jason-2 only results in Figure 2.5 are less marked. Variation for all data sources is of the order of 1%.

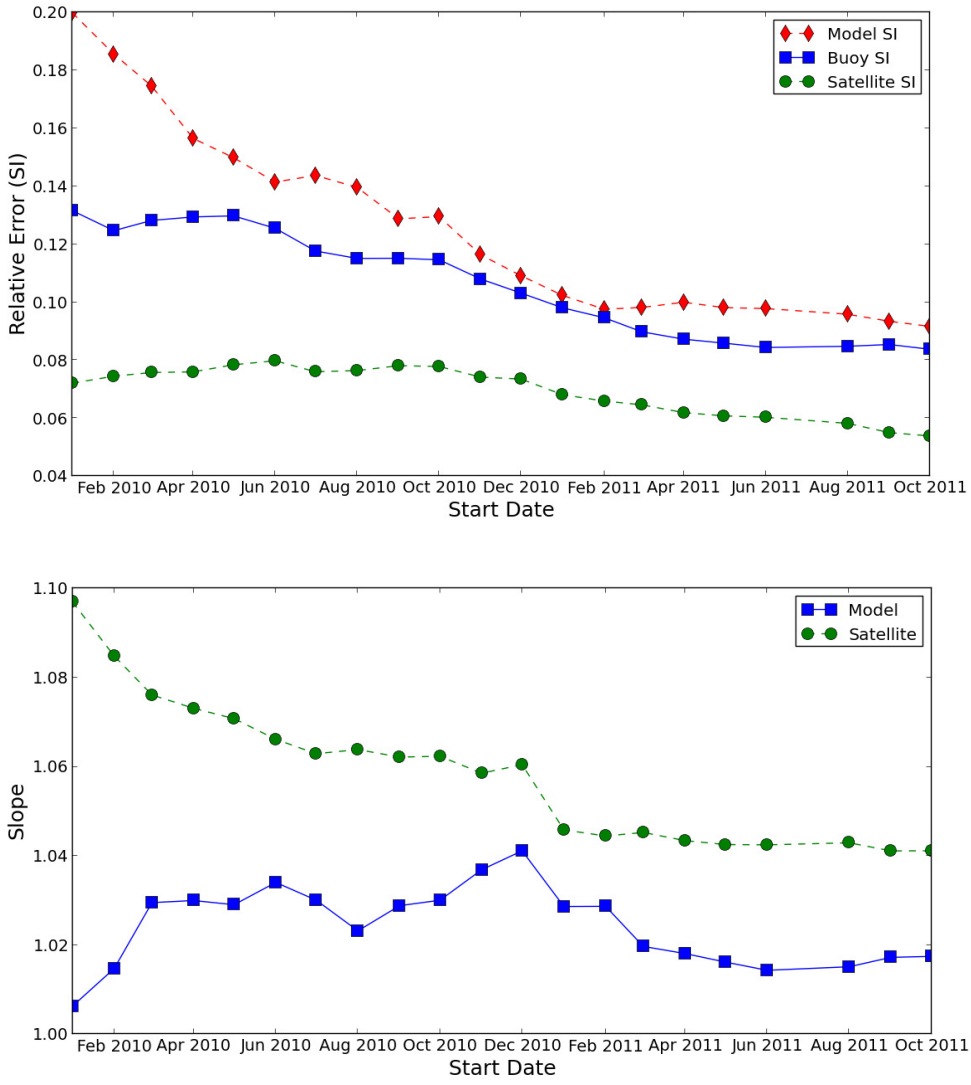


Figure 2.4. Time-series of rolling 12 month sample error estimates for model, in-situ platform (labelled buoy) and satellite SI (upper panel) and slope (lower panel) for the North Sea. Samples were determined based on Jason-2 altimeter data with 3 soundings in each super-observation; the in-situ data provided the reference 'truth' for the slope estimate.

II.3.2 Sensitivity to use of independent matchup data

A further alteration to the matchup sample can be made by choosing to strictly enforce independence criteria in the data. These tests were applied for the North Sea matchups, which (as illustrated in Figure 2.5) include several clusters of closely located in-situ platforms. Since the use of subsampling reduces the sample size available, the independence

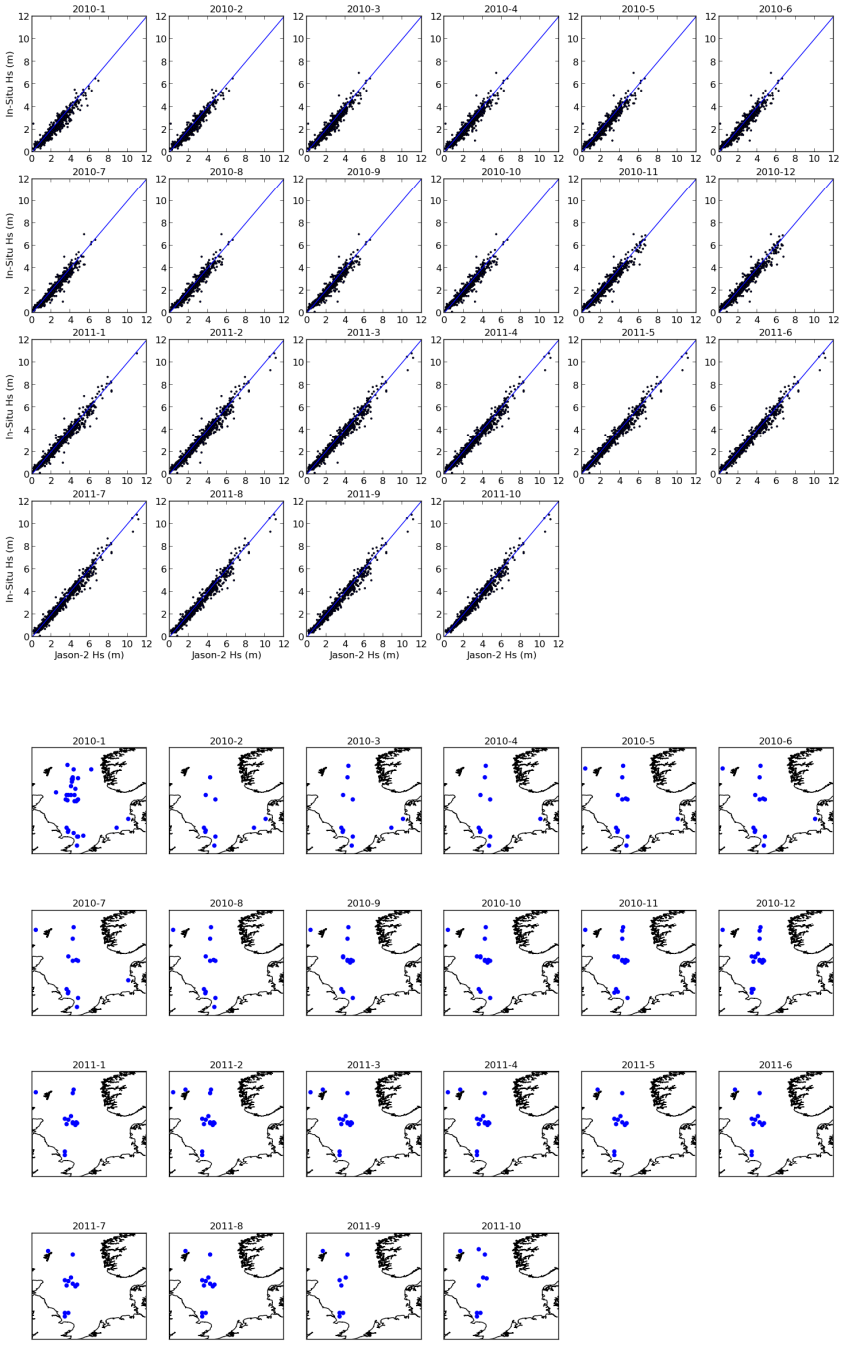


Figure 2.5. Hs scatter data (Jason-2 vs in-situ, top panel) and unique in-situ locations used in rolling 12 month triple collocation samples for North Sea analyses.

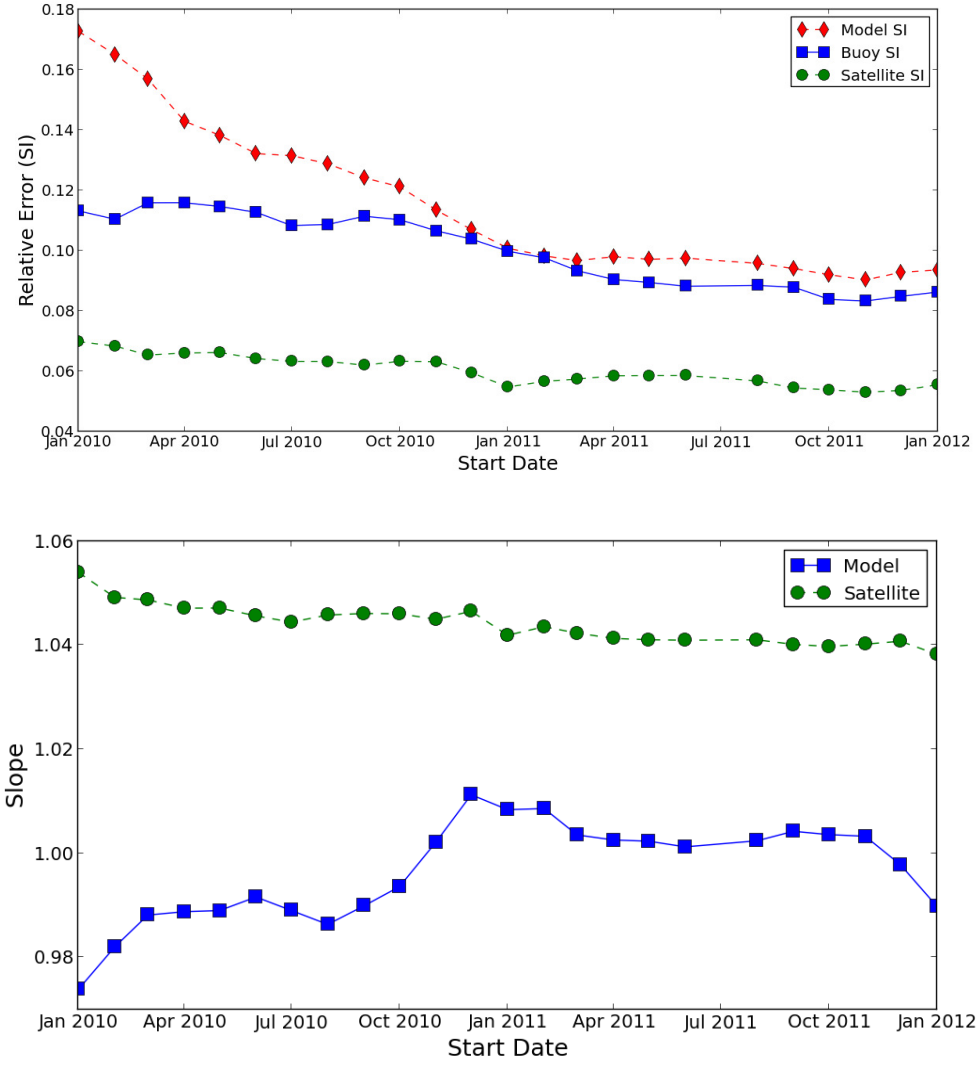


Figure 2.6. Time-series of rolling 12 month sample error estimates for model, in-situ platform (labelled buoy) and satellite SI (upper panel) and slope (lower panel) for NEAM after removal of Brittany outlier data. Samples were determined based on all available Envisat, Jason-1 and Jason-2 altimeter data with 3 soundings in each super-observation; the in-situ data provided the reference 'truth' for the slope estimate.

sensitivity tests were carried out on the full period available for analysis. Table 2.5 shows the standard deviation of SI and slope estimates taken from analyses of an ensemble of 1000 matchup subsamples Each subsample was derived by randomly drawing unique independent data based on the time and distance criteria given in subsection II.2.3. The

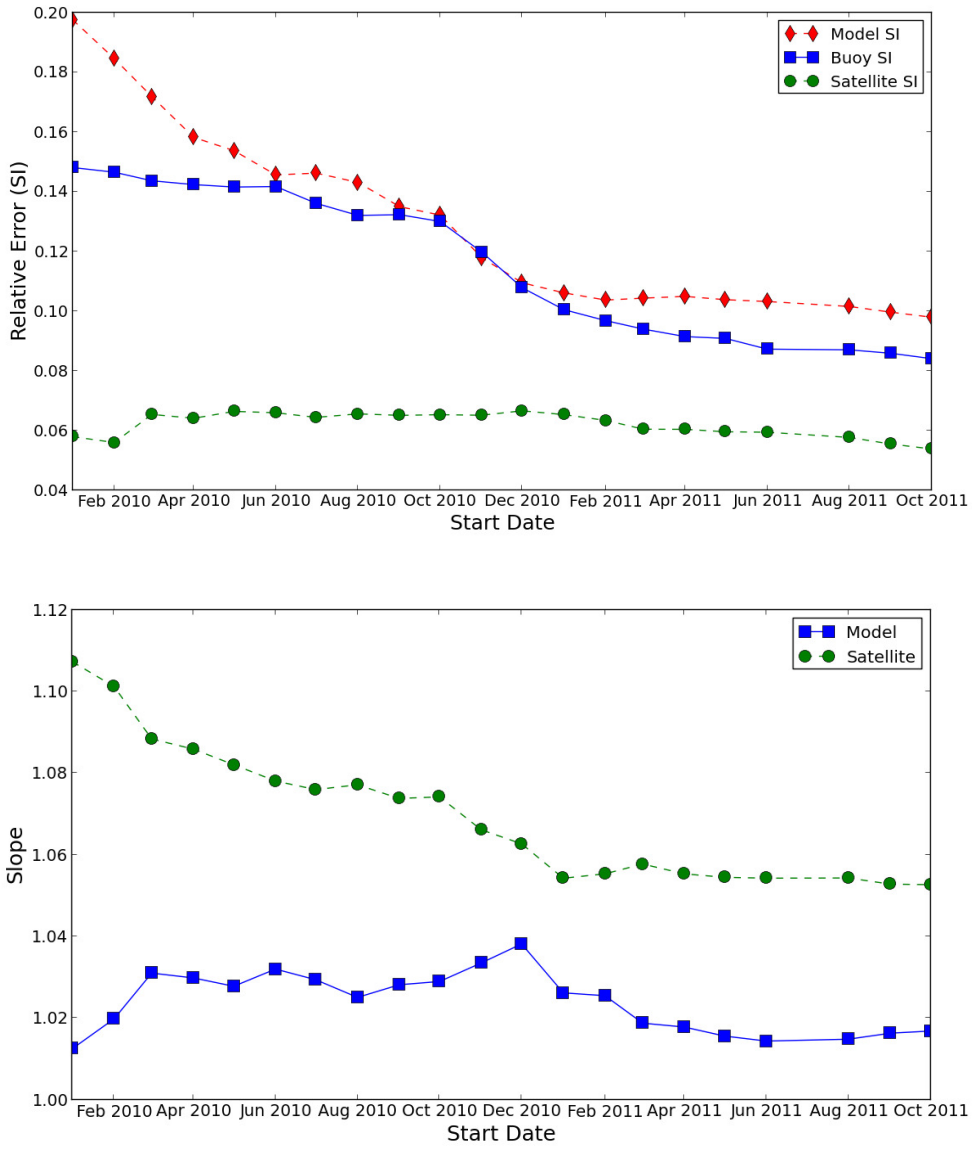


Figure 2.7. Time-series of rolling 12 month sample error estimates for model, in-situ platform (labelled buoy) and satellite SI (upper panel) and slope (lower panel) for the North Sea. Samples were determined based on all available Envisat, Jason-1 and Jason-2 altimeter data with 3 soundings in each super-observation; the in-situ data provided the reference 'truth' for the slope estimate.

impact of this subsampling was virtually negligible; the mean SI and slope values did not change from an analysis of the full dataset by more than 0.1% and the standard deviation from the mean was also less than 0.05%.

Table 2.5. Distance between observation independence criteria subsampling results for North Sea error estimates; standard deviation of estimates from 1000 member ensemble of random draws from a match-up sample determined based on Jason-2 altimeter data with 3 soundings in each super-observation; the in-situ data provided the reference 'truth' for the slope estimate.

Distance	Model SI	Buoy SI	Satellite SI	Model Slope	Satellite Slope
100	0.00037	0.000036	0.00037	0.00042	0.00040
150	0.00018	0.00014	0.00018	0.00020	0.00019

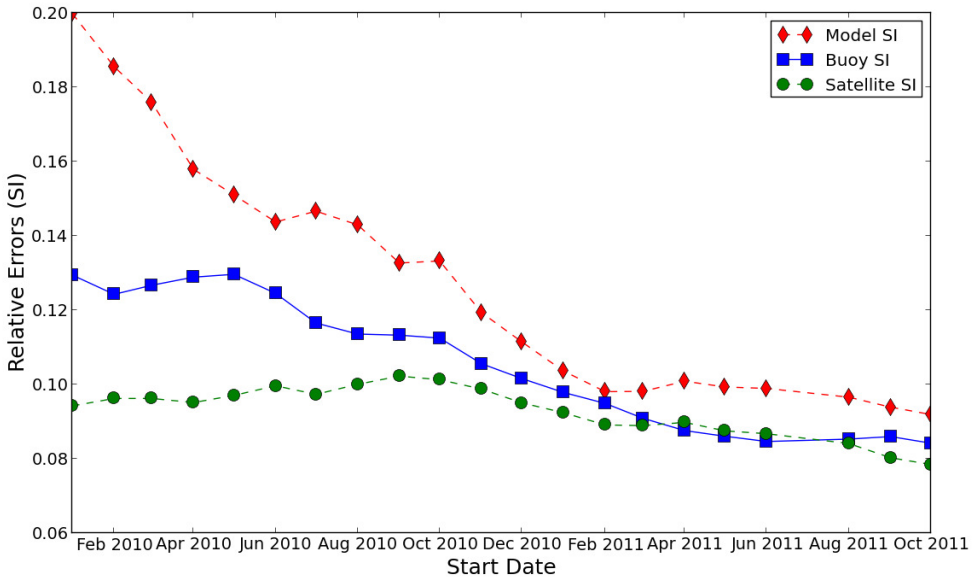
II.3.3 Effects of satellite super-observation

The dependence of the error estimates on satellite super-observation were tested by comparing results derived when 1, 3, and 5 averaged altimeter soundings were used in generating the satellite data used in the matchup samples. Results of these tests showed that variability in the error estimates were isolated purely to the satellite and its SI. Here the effect of super-observation was to reduce SI values by approximately 1-2% with each successive super-observation step, as is illustrated for the Jason-2 case in the North Sea in Figure 2.8. The results for 1 sounding and 5 sounding super-observation can be compared directly with the 3 sounding super-observation data in the upper panel in Figure 2.4. Whilst it might be expected that there will be an asymptotic limit to how low the error can become with further super-observation this was not tested and so, for the scales one might wish the observations to represent for the purpose of regional verification (equivalent to approximately 20km, see MyWave-WP4.2a), the choice of satellite super-observation methodology is expected to be somewhat subjective.

II.3.4 Summary of overall results

Based on the sensitivities noted in the previous subsections an overall analysis of the data from period 2010-2012 was made based on matchup samples using 3 sounding altimeter super-observation and no subsampling criteria. In view of sample limitations and changes with time however, bootstrap resampling (Efron and Gong, 1983) was applied to the data to create a 1000 member ensemble of matchup data from which the effects of sample variability could be judged.

Single altimeter sounding



5 sounding super-observation

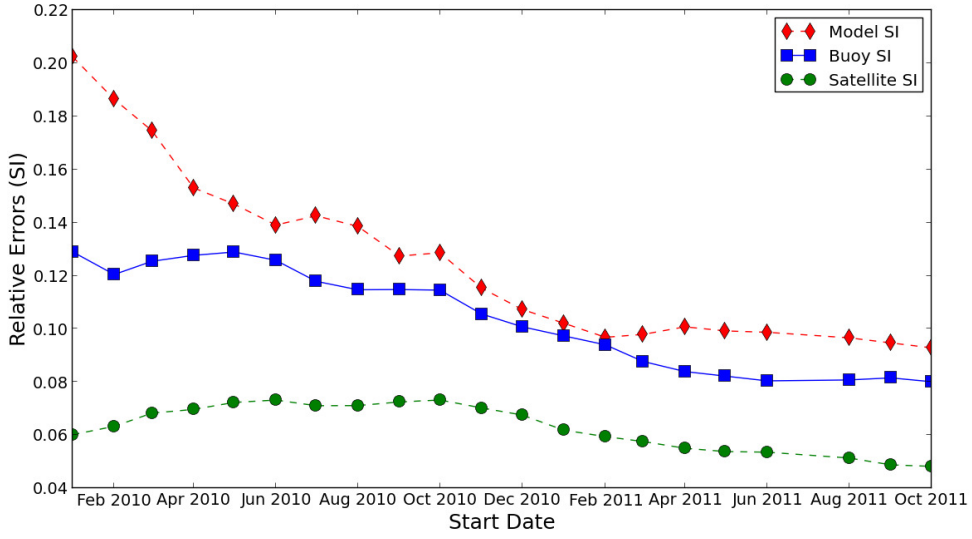


Figure 2.8. Time-series of rolling 12 month sample error estimates for model, in-situ platform (labelled buoy) and satellite SI for the North Sea. Samples were determined based on Jason-2 altimeter data with (upper panel) 1 sounding and (lower panel) 5 soundings in each super-observation.

Table 2.6. Bootstrap ensemble mean SI and slope values for 2010-2012 period in the NEAM; for matchup samples determined using Jason-2 data only and all satellites.

Satellite	Model SI	Buoy SI	Satellite SI	Model Slope	Satellite Slope
Jason-2	0.125	0.106	0.055	0.969	1.024
All	0.126	0.114	0.063	0.996	1.046

Table 2.7. Bootstrap ensemble standard deviation of SI and slope values for 2010-2012 period in the NEAM; for matchup samples determined using Jason-2 data only and all satellites.

Satellite	Model SI	Buoy SI	Satellite SI	Model Slope	Satellite Slope
Jason-2	0.0036	0.0078	0.0048	0.0043	0.0033
All	0.0025	0.0085	0.0031	0.003	0.0026

Table 2.8. Bootstrap ensemble mean SI and slope values for 2010-2012 period in the North Sea; for matchup samples determined using Jason-2 data only and all satellites.

Satellite	Model SI	Buoy SI	Satellite SI	Model Slope	Satellite Slope
Jason-2	0.133	0.105	0.072	1.019	1.058
All	0.130	0.110	0.065	1.020	1.066

Table 2.9. Bootstrap ensemble standard deviation of SI and slope values for 2010-2012 period in the North Sea; for matchup samples determined using Jason-2 data only and all satellites.

Satellite	Model SI	Buoy SI	Satellite SI	Model Slope	Satellite Slope
Jason-2	0.0035	0.0041	0.0038	0.0042	0.0034
All data	0.0025	0.0054	0.0031	0.0032	0.0028

Results are given in Tables 2.6 and 2.7 for the NEAM and Tables 2.8 and 2.9 for the North Sea, and are based on overall matchup samples and those using only the Jason-2 data. Over the two regions model relative errors (SI) were reasonably consistent (within 1-2%) and an overall estimate from the test period is approximately 12–13%, in-situ SI is 10–11% and the satellite SI was lowest of all at approximately 5–7%. Systematic errors (slope) were more strongly differentiated between the two regions (by approximately 3-4%), with the in-situ data seen to be lower relative to the other data sources in the NEAM. The comparison between Jason-2 only data and the 'all satellite' sample was differentiated by less than 1%. Calculated standard deviations from the bootstrap ensemble fell between 0.2-0.5% for all error estimates, suggesting that a good confidence can be placed on the values provided in this analysis within +/-1% of the those estimated by the bootstrap mean.

II.4 Discussion

This study has demonstrated that it is feasible to generate sensible triple collocation estimates of observation errors on a regional basis following the JEA07 method. In general, results for relative error (SI) and linear calibration coefficient (slope) were found to be accurate within +/-1%. Key sensitivities in the results were to changes in the in-situ data used and, for satellite SI only, changes in super-observation. Within the regions analysed differentials in relative error shown by in-situ and satellite based observing systems were generally limited, however the differentials in linear calibration constant appear sufficiently different to justify a regional approach.

No major requirement to ensure location-time independence in the underlying matchup data was found in the samples analysed, which is useful since this allows the number of data used in the matchup sample to be maximised and this in turn improves the likelihood that the error estimates will be convergent and robust. However, the results can be particularly sensitive to outliers and, with only a small number of in-situ data available in most regional sea areas, any changes in the in-situ network. Therefore significant effort may need to be spent in quality controlling data in an operational triple collocation assessment that is sample constrained in space and time. On the other hand, the technique was proven very adept in tracking background changes in the data sources assessed and provided a stable estimate of observation errors even when the model data used was a 'moving target'.

III APPLICATION OF OBSERVATION ERRORS TO VERIFICATION METRICS

III.1 User requirement

This section describes a method by which observation error data, for example as calculated using the triple colocation technique described in the previous section, can be used to assist users understand and contextualise the verification data.

MyWave report D4.1 noted various analytical and simulation methods with which metric data might be corrected to include the effects of observation errors (e.g. following Tolman, 1998; Saetra and Bidlot, 2004; Saetra et al., 2004; Bowler et al., 2006). However, through consultation with users WP4 has identified a preference for users to be presented with verification data that focuses on a direct comparison between prediction and reference observation (rather than a corrected result); which retains a separation of metrics measured against different observed references (e.g. in-situ and satellite data); which quantifies the verification in real terms rather than as a skill score; and which aims to contextualise the results of the metrics where possible (see MyWave-D4.2a).

This feedback has focused the project toward methods that contrast direct comparisons between prediction and reference with an 'idealised' verification that quantifies the effect of observation errors.

III.2 Method to contextualise verification metric data

In the proposed method three steps are taken to contextualise the direct comparison verification results:

1. Re-sampling to help understand the effects of the sample used.
2. Generation of idealised verification data to estimate target performance levels.
3. Generation of naïve prediction verification data to define a low performance boundary.

The effect of the sample used in verification is tested through re-sampling the prediction-reference matchup data, calculating an individual metric many times and showing the range of results this produces. The use of re-sampling is particularly beneficial when dealing with small sample sizes that may be aliased by outlying data. Making the calculations many times also enables a robust assessment of the effect of observation errors when estimating the idealised verification.

The underlying assumption for the idealised verification is that the original prediction data makes a reasonable estimate of the observed climate. If this holds a metric can then be calculated using the scenario where the prediction correctly represents the truth and the observation is represented by prediction plus a draw from an observation error probability distribution function (pdf). In this case errors are attributed to the observation and the resulting scores estimate how well the truth might expect to perform against the observation as a reference. Since the observations are simulated the effect of the observation errors can be tested for a large variety of metrics. The simulations are highly unlikely to replicate the exact observation errors incorporated into the direct prediction-reference comparison, but applying the error simulation multiple times within a re-sampled ensemble should generate a robust dataset and enable easy comparison with direct prediction-observation based metrics.

In consultation users were opposed to working with a skill score that combined direct comparison data with a naïve prediction. However, when treated separately in order to define a low performance boundary, it is believed that the evaluation of the performance of an unskilled naive prediction lends further context (this position will be made subject to further testing with users as the project progresses). Numerous options are available for the naive prediction, but in this document we propose the use of a random draw from the original sample of prediction data. This is suggested in preference to observation derived baselines since the naïve prediction retains systematic bias features of the originating prediction system. Besides, the prediction system is logically the only source from which forecast data can be provided apriori to the observed events.

The use of idealised and naïve prediction metric data allows the direct prediction-observation metric to be placed within a continuum as indicated in Figure 3.1. The optimum score for the metric is achieved when the prediction agrees perfectly with the reference data. However, the idealised scenario verification will fall some distance away from the optimal score as a result of the observation errors and the aim of the direct comparison should be to fall close to

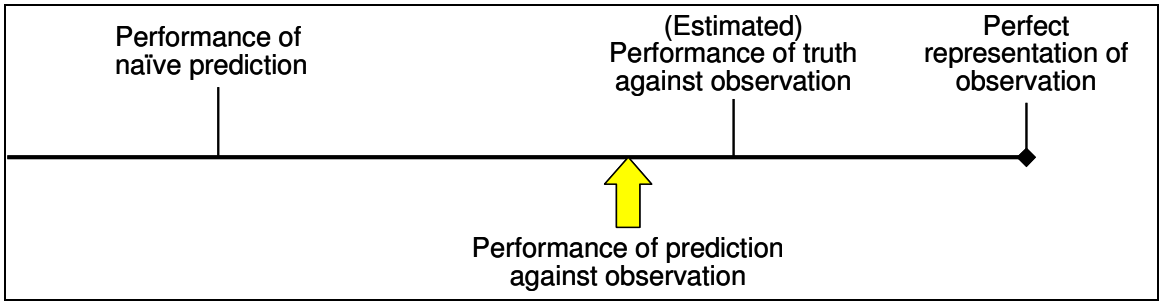


Figure 3.1. Schematic of a continuum for metric values.

this value. The naïve prediction metric provides a baseline for low levels of performance and the value of the prediction system being verified is questionable when close to this score.

III.3 Example application

This section describes an example application of the principles described in subsection 3.2 to verification data. The verification sample used compares predictions from the Met Office global (35km) wave model against in-situ data taken from the JCOMM international intercomparison of operational ocean wave forecasting systems maintained by ECMWF (Bidlot et al., 2007) and satellite observations derived from Jason-2 near real-time altimeter data. In the following examples the performance of the model is assessed for a region of the North Sea (Figure 3.2) during the months from January-March 2012.

III.3.1Block bootstrap data sampling

The block bootstrap (Carlstein, 1986; Kunsch, 1989) approach in this application uses data acquired from the spatial blocks indicated in Figure 3.2 and with the temporal block time set at 24 hours. In order for the blocks to carry equal weight in the verification, a standard sample size is set and data up to this size are drawn randomly (without replacement) from the block sample each time the block is accessed. For the in-situ data this led to a sample size of approximately 2200 data points distributed over 270 blocks and for the Jason-2 data approximately 300 data points distributed over 60 blocks. The blocks are accessed and assigned to the verification sample using a standard bootstrap draw with replacement. In this case an ensemble comprising 1000 bootstrap members was generated.

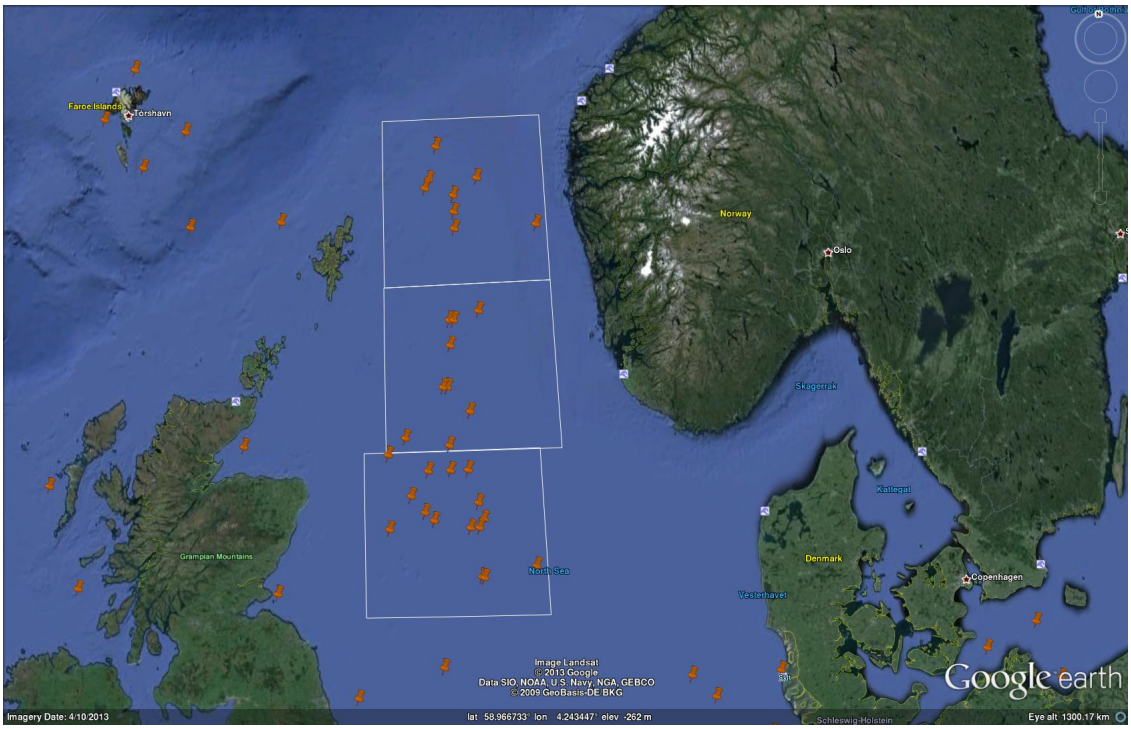


Figure 3.2. Positions of in-situ observation platforms and area blocks (white squares) for the block bootstrap applied to Central-Northern North Sea area verification.

III.3.2 Generation of idealised observation data

For each bootstrap member a pseudo-observation was generated by applying a simulation of the observation errors to the original prediction member. The form of the errors affect certain metrics differently, for example an idealised (unbiased) RMSE estimate will depend primarily on the scale value for the assumed observation error pdf, whilst a stratification by predicted quantity will yield results that have a dependency on whether the errors are assumed to be homo- or heteroscedastic (e.g. Figures 3.8-3.9). In the examples provided the assumed error distribution is a standard normal and heteroscedasticity is expected following the logic that the largest observation errors are more likely to occur when measuring the highest energy conditions. The model used then takes the form:

$$OE_i = \beta M_i + \mathrm{N}(0, \sigma),$$

where OE are the idealised observations, M are the prediction data, β beta is the observation slope data and the normal distribution scale factor σ is calculated as the multiple of the mean prediction value with the observation error scatter index (SI) as described by JEA07. The

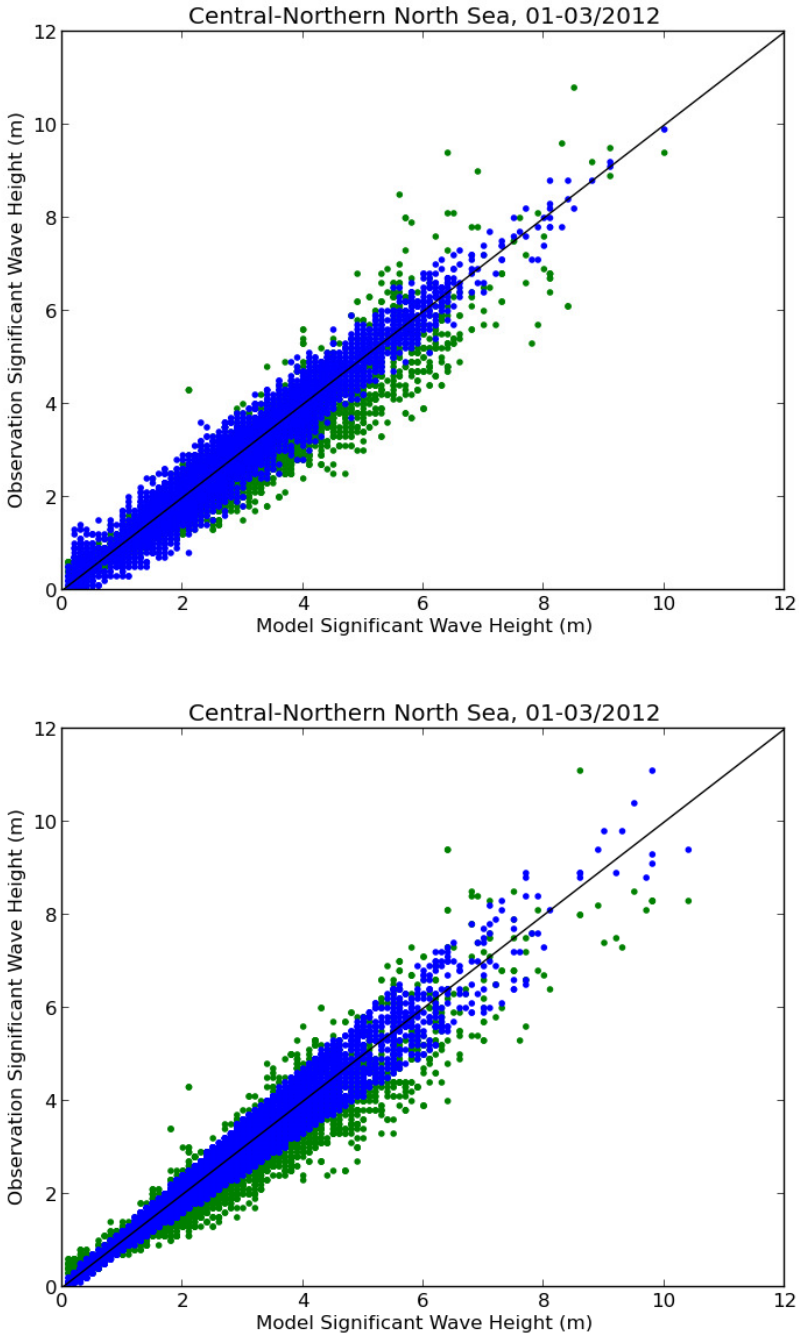


Figure 3.3. Comparison of significant wave height scatter data for (upper panel) model versus in-situ observed data (green) and an idealised observation generated from model using a homoscedastic assumption (blue); (lower panel) model versus in-situ observed data (green) and an idealised observation generated from model using a heteroscedastic assumption (blue).

effect of using the heteroscedastic assumption is shown in Figure 3.3. For the simulations a value of β = 1.0, SI = 10% were used for the in-situ data and β = 1.05, SI = 6% were used for the Jason-2 data based on results in Section II.

III.3.3 Naïve prediction

For each ensemble member the naïve predictions were generated by randomising the order of the original prediction sample.

III.3.4 Application to metrics

The metrics shown in this subsection are a subset of metrics described in MyWave-D.4.2a. Examples presented are:

- Test C2: Quantile-quantile plot
- Test M1: Symmetrically Normalised Root Mean Square Error (SNRMSE; Mentaschi et al., 2013)
- Test P1: Probability of significant wave height (Hs) predicted within 0.25m of the reference value
- Test P2: Success ratio for prediction of Hs greater than 2m
- Test R2: Bias and error standard deviation through Hs prediction range.

Figure 3.4 illustrates application of the method to a quantile-quantile plot. In the figure the blue 'point and cross' symbols show the direct comparison between model and observation and the green symbols show the idealised scenario. The point values are sited at the location of the bootstrap ensemble mean values (in this case Hs for every second percentile from 2-98% of the distribution) and the extent of the crosses represent 5% and 95% values across the ensemble. In the upper panel of Figure 3.4, the idealised data are derived using in-situ error estimates (which are assumed unbiased) and so the ideal case lies almost directly on the 1:1 line in the plot. In the lower panel the comparison is made against Jason-2 data and the idealised data can be seen to be biased away from the 1:1 line toward the reference (y-axis) as a result of the slope correction. The cross-lengths are significantly larger for the satellite verification than for the in-situ verification illustrating the uncertainties associated with the low sample size in the Jason-2 matchup dataset. The comparison between model and observation is illustrated by the deviation from the 1:1 line, and the

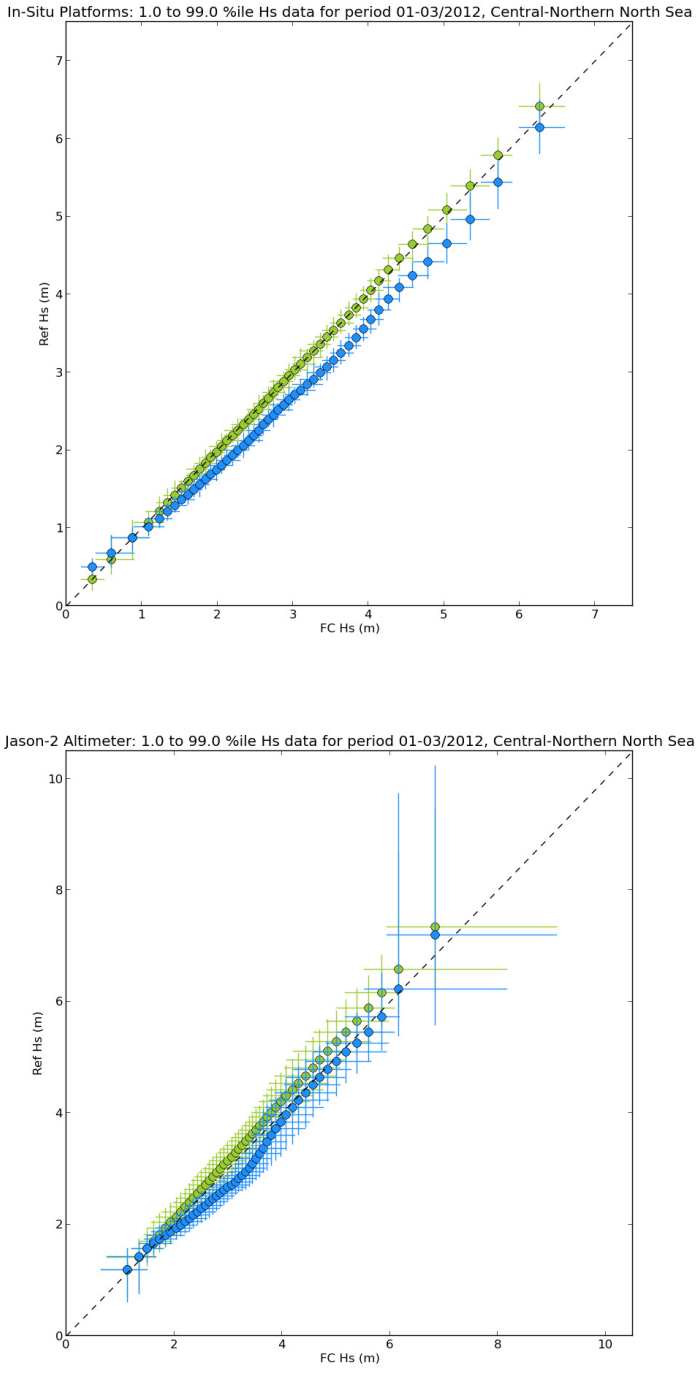


Figure 3.4. Quantile-quantile plots for (upper panel) model (at T+0) versus in-situ observations (blue) and idealised observation (green); (lower panel) model (at T+0) versus Jason-2 observations (blue) and idealised observation (green). Quantiles are shown every 2% from 2-98%, marker centres show bootstrap ensemble mean and cross extents show ensemble 5% and 95% values.

comparison between model and estimated truth is illustrated by the deviation between the blue and green symbols. In both cases, above 2m, there is insufficient overlap in the cross part of the symbols to imply that differences in the data are an effect of sample size.

Figure 3.5 illustrates the application of the method to SNRMSE, which for the model is compared through lead time. Context is given by plotting the idealised metric value (green) and naïve prediction (orange) scores associated with the T+0 model data. For this metric the idealised data are derived directly from the RMS of the observation error distribution and are related to the applied slope and scale factors. The T+0 context data are applied constantly across the plot in this instance since we assume that at T+0 the model makes its best representation of climatology (used in generating the idealised data) and that a prediction using a random ordering of the model climate would be the best available naïve prediction regardless of lead time. For the idealised and naïve metric scores the variation introduced within the bootstrap ensemble is shown using a banded plume (set at 1%, 5% and 25% from the distribution tails). For the direct model-observation comparison a box and whiskers display is used. In this instance the centre value represents the ensemble mean, the box outer and inner lines indicate 5% and 25% data values from the distribution tails and the flyers indicate the 1% quantiles from the tails. Interpreting the plot, a number of conclusions can be drawn. The direct comparison data show sufficient change with increasing lead time versus spread in the box and whiskers to expect that the increase in SNRMSE with lead time shown in the plot is genuine. The direct comparison errors are substantially higher than the idealised scenario, suggesting that there is room for improvement in the model, but the direct comparison errors are in turn much smaller than the naïve prediction errors, suggesting that overall the model is skilful even at the 5 day range. Spread in both the naïve prediction and T+120 SNRMSE ensembles are high, indicating that the value taken by the metric in lower skill situations is significantly affected by sample size and effects of random chance within the match-up sample. In contrast the target idealised values are very stable.

Figure 3.6 shows application to a metric testing probability of the prediction falling within 0.25m of the observed value. The plot is set out in the same manner as for Figure 3.5. In this case the metric has a strong dependency on the background wave climate, and since the waves are relatively high at this time of year even the idealised case achieves a relatively low score (which is a function of the threshold used and both scale factor and pdf applied to the observation error distribution). Overall, variability is higher in this plot than in Figure 3.5,

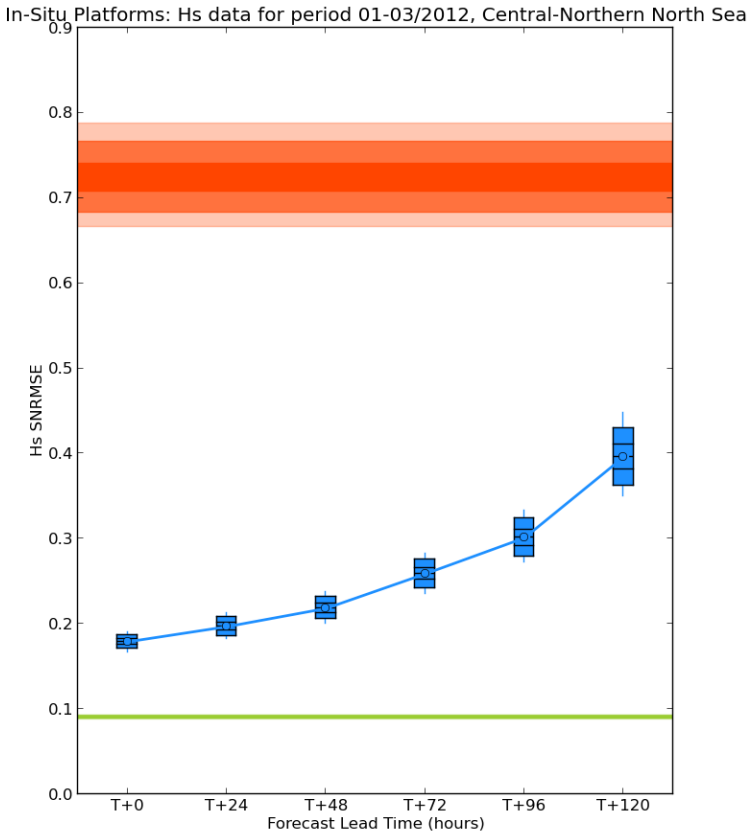


Figure 3.5. Symmetrically Normalised Root Mean Square Error (SNRMSE) versus lead time for model against in-situ platform data. Box and whiskers symbols show the direct model-observation comparison (marker at bootstrap ensemble mean, inner box lines at 25-75% range, outer box lines at 5-95% range and flyers at 1-99% range), the green plume shows idealised SNRMSE (same ranges) and the orange plume shows the naïve prediction SNRMSE (same ranges).

partially due to the use of Jason-2 data as the reference leading to a lower sample size. Variability of the idealised case is higher than for the naïve prediction, indicating that the sample affects this metric substantially in terms of identifying moderate or good performance and that a well defined lower limit for performance exists for this metric. The influence of the sample on the direct comparison is relatively constant through the forecast range but the trend for decreasing performance with increasing lead time is still clear, as is the potential for the model to improve further in terms of this metric at all lead times. Indeed, at T+120 the

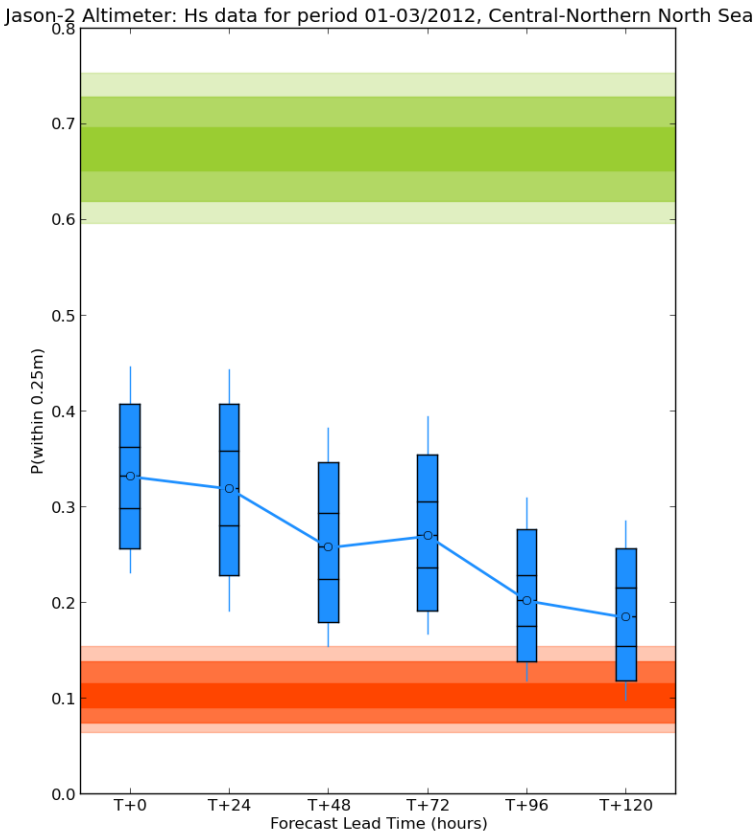


Figure 3.6. Probability of prediction falling within 0.25m of reference value versus lead time for model against Jason-2 data. Box and whiskers symbols show the direct model-observation comparison (marker at bootstrap ensemble mean, inner box lines at 25-75% range, outer box lines at 5-95% range and flyers at 1-99% range), the green plume shows idealised SNRMSE (same ranges) and the orange plume shows the naïve prediction SNRMSE (same ranges).

model might be expected to have limited skill in meeting this particular criterion when compared with satellite data.

Figure 3.7 shows a success ratio metric, which will be heavily influenced by the background climate in terms of the proportion of the sample that contributes to the score. In this case a relatively high number of forecasts are well above the 2m threshold and all the predictions score highly on the metric. However, by contextualising the data it can be seen that although

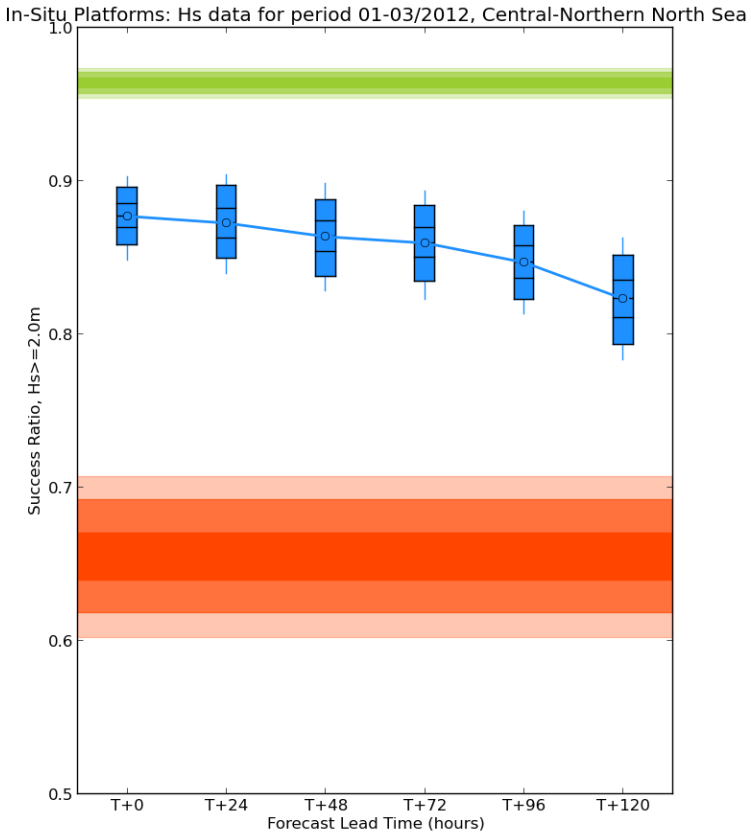


Figure 3.7. Success Ratio for forecasts of Hs greater than 2m versus lead time for model against in-situ data. Box and whiskers symbols show the direct model-observation comparison (marker at bootstrap ensemble mean, inner box lines at 25-75% range, outer box lines at 5-95% range and flyers at 1-99% range), the green plume shows idealised SNRMSE (same ranges) and the orange plume shows the naïve prediction SNRMSE (same ranges).

the model could improve throughout the forecast range and the predictions are reasonably skilful against the in-situ observations even at T+120.

Figure 3.8 examines performance of the model, in terms of bias and error standard deviation, through the predicted significant wave height range (for a lead time of 48 hours). In this case the data are stratified in (overlapping) 20% quantile ranges of the predicted significant wave height in order that sample size is consistent. If the stratification method used leads to very

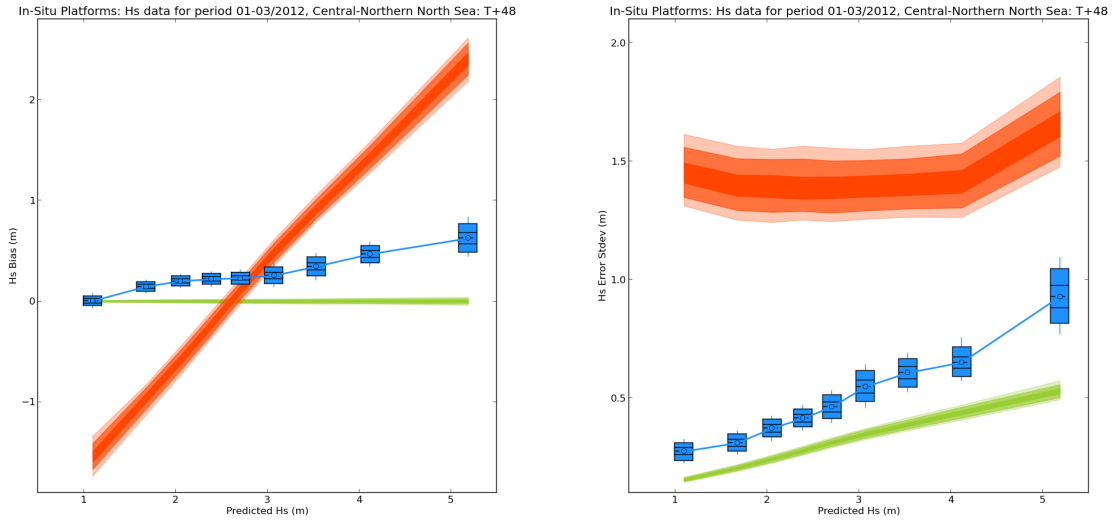


Figure 3.8. (left panel) Bias and (right panel) Error Standard Deviation versus predicted significant wave height for model against in-situ data. Box and whiskers symbols show the direct model-observation comparison (marker at bootstrap ensemble mean, inner box lines at 25-75% range, outer box lines at 5-95% range and flyers at 1-99% range), the green plume shows idealised SNRMSE (same ranges) and the orange plume shows the naïve prediction SNRMSE (same ranges).

small samples then direct comparison between the metric and an analytical solution (which assumes a sample size that converges the metric is achieved) might be misleading. Therefore the simulation technique used here enables a more robust comparison.

In the bias plot the naïve prediction shows large differentials in bias from underprediction when wave heights are low, to overprediction when wave heights are high. This is to be expected as the random ordering of the prediction data should mean that the average value in each prediction bin is compared with the full sample mean for each bias value calculated. The direct comparison data show a similar, if much less marked, drift with increasing wave height (which again should be a feature if the prediction and observation are not fully correlated), whilst the idealised case shows only a small (sample based) variability around zero. As might be expected, standard deviation of the errors for the direct comparison increase markedly for predictions in the tail of the distribution (since in the tail the chance of encountering a significantly different value in the observations is proportionately higher than

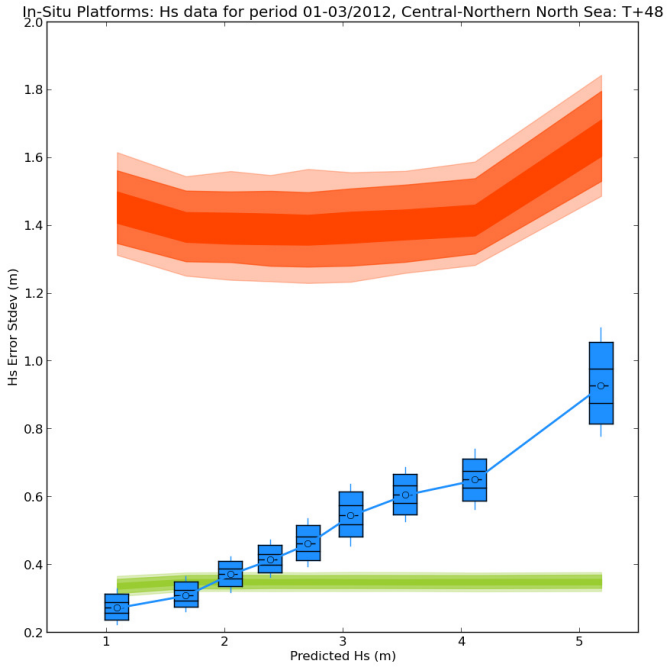


Figure 3.9. Error Standard Deviation versus predicted wave height for model against in-situ data using a homoscedastic error model. Box and whiskers symbols show the direct model-observation comparison (marker at bootstrap ensemble mean, inner box lines at 25-75% range, outer box lines at 5-95% range and flyers at 1-99% range), the green plume shows idealised SNRMSE (same ranges) and the orange plume shows the naïve prediction SNRMSE (same ranges).

for predictions from the main body of the wave height distribution). The comparison with the naïve prediction data suggests that whilst high value predictions are quantitatively less skilful than low-mid value predictions, there is still predictive ability in the model at T+48. This conclusion is emphasised by the slope that can be seen in the idealised prediction data where, as a result of the heteroscedastic assumption, the idealised error standard deviation also increases with predicted wave height. For comparison Figure 3.9 shows the same analysis based on a homoscedastic error distribution, which shows a flatter profile to the idealised data plume and a scenario where the direct error comparison falls below the idealised case for low predicted wave heights. This suggests some flaws in an error model that assumes the errors are entirely independent from the background wave height.

III.4 Comparison of metrics derived against different baselines

The technique discussed can be applied generically across a number of metrics and for comparisons against both in-situ and satellite based observations. In order to compare data derived against different references the idea of contextualising the metric results against idealised and naïve prediction results is taken further. This process necessitates generating a skill score since the comparison must in some way account for the fact that (as demonstrated in MyWave-D4.2a) the background data sampled against different references will vary in terms of its climatology and can in turn affect absolute metric values.

To make the comparison a normalisation is employed that measures the direct comparison value against the 'skill gap' between naïve prediction and perfect replication of the observation (Figure 3.1):

$$OPS = 1.0 - \frac{V_M - OV}{V_N - OV},$$

where OPS is defined the 'observation prediction skill' of the system (above model background), OV is the optimal verification score (perfect replication of observations), V_M is the verification score for the model and V_N is the verification score for the naïve prediction.

The justification for using this scale is threefold. Firstly these comparisons are intentionally differentiated in purpose from the quantified metric comparisons shown in subsection 3.3, and are expected to be mainly of interest to model developers attempting to identify where the model adds skill over and above background. Secondly, the naïve prediction performance provides a less subjective benchmark than the idealised data, since the latter have some dependency on the form of error model chosen. Last, and crucially, the normalising factor (a function of the naïve prediction score) will include background condition effects. To illustrate this a simple case can be considered where the naïve prediction and observations can be represented by independent normally distributed variables. In this instance the errors generated when comparing the two samples will also be a normal distribution with a location value equal to the systematic error between the two samples and a scale parameter which is a function of the sum of the variance associated with each sample. As a result the error distribution will increase in variability in line with an increase in the background variability of the observations and naïve prediction.

III.4.1 Example application

Figures 3.10-3.12 examine properties of this comparison by examining the variation of the observation prediction skill for different metrics derived from a rolling 12 month collection of 3 month data samples. These are subject to significant variability in terms of background wave conditions, with mean Hs varying from 2.4-2.7m in the winter samples to 1.4-1.8m in the summer samples. In the figures the months are labelled using the first of the 3 months in each sample (e.g. 1201 corresponds to January-March 2012).

In Figure 3.10 the statistic tested is MAE, which shows significant seasonal variability through the period examined (upper panel). The variability has a clear link to changes in background conditions. Whilst the changes in the metric value through time will have a connection to varying model skill in predicting different backgrounds, the metric may also be influenced more directly by what is being observed. By applying the comparative measure some of this variability is reduced (lower panel) and the clearest change in prediction skill occurs toward the end of 2012. Figure 3.11 shows the same comparison applied to probability of the prediction falling within 0.25m of the reference. The metric results show very strong seasonal variability, with an improvement in both the metric data and the observation prediction skill during the more benign spring and summer months. Figure 3.12 shows the differential in observation prediction skill between direct comparison and the idealised case. When this comparison is made, although the quantitative values differ, a similar performance pattern emerges for both metrics with the optimal performance in the model shown for the spring-summer transition and in the autumn-winter data in late 2012 (post a supercomputing upgrade at the Met Office). A similar comparison for success ratio for prediction of Hs greater than 2m (not shown) provides the same qualitative conclusions.

Examining the metrics in this manner suggests that the use of an observation prediction skill normalisation de-sensitises the metric to background conditions to some degree, but retains certain variability characteristics. Comparing the differential to the idealised case, although more subjective, may offer some ability to compare performance between different metrics. However this would need testing on a larger set of metrics than has been attempted here.

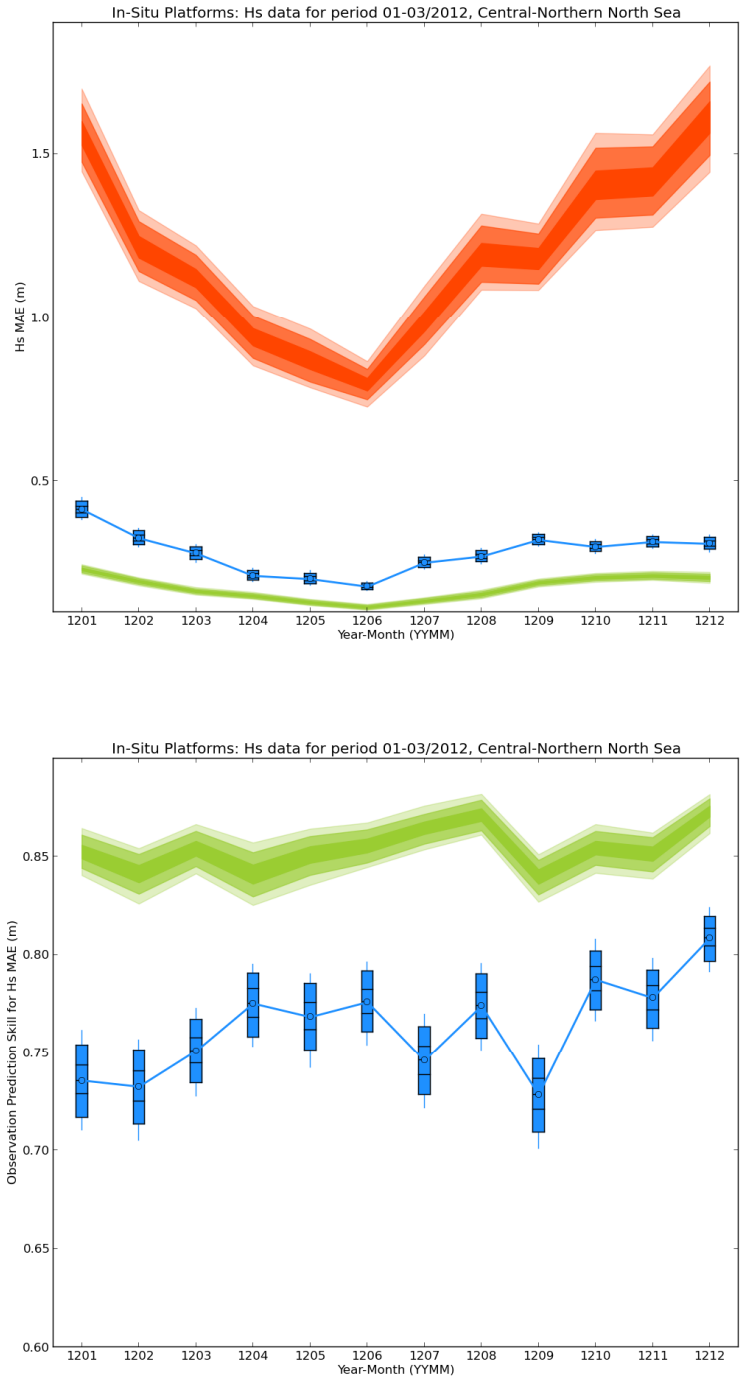


Figure 3.10. Mean Absolute Error (MAE) for rolling 3 month samples during 2012 using model (T+0) against in-situ data. (upper panel) Direct comparison, idealised prediction and naïve prediction verification displayed using the same schema as for Figures 3.4-3.9; (lower panel) MAE Observation prediction skill scores for direct comparison (box and whiskers data) and idealised prediction.

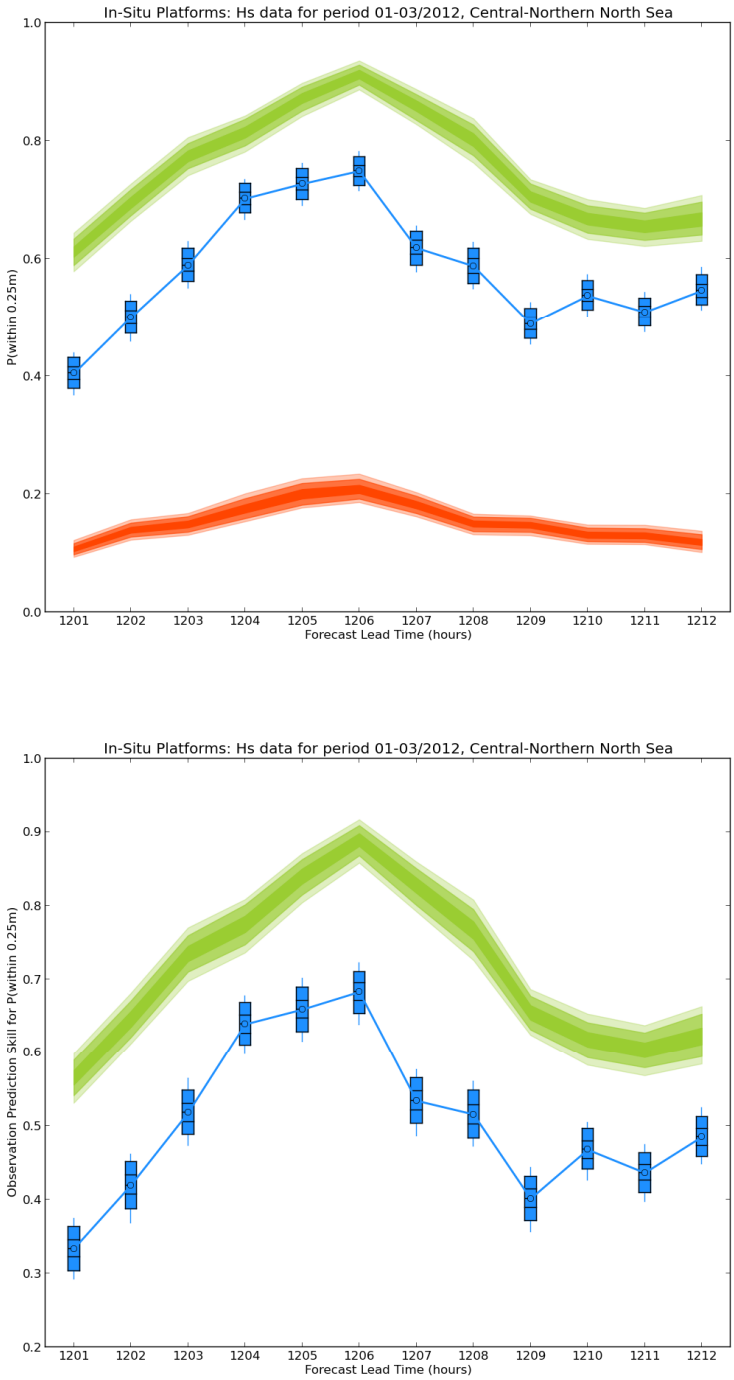


Figure 3.11. Probability of prediction within 0.25m of reference for rolling 3 month samples during 2012 using model (T+0) against in-situ data. (upper panel) Direct comparison, idealised prediction and naïve prediction verification displayed using the same schema as for Figures 3.4-3.9; (lower panel) Observation prediction skill scores for direct comparison (box and whiskers data) and idealised prediction.

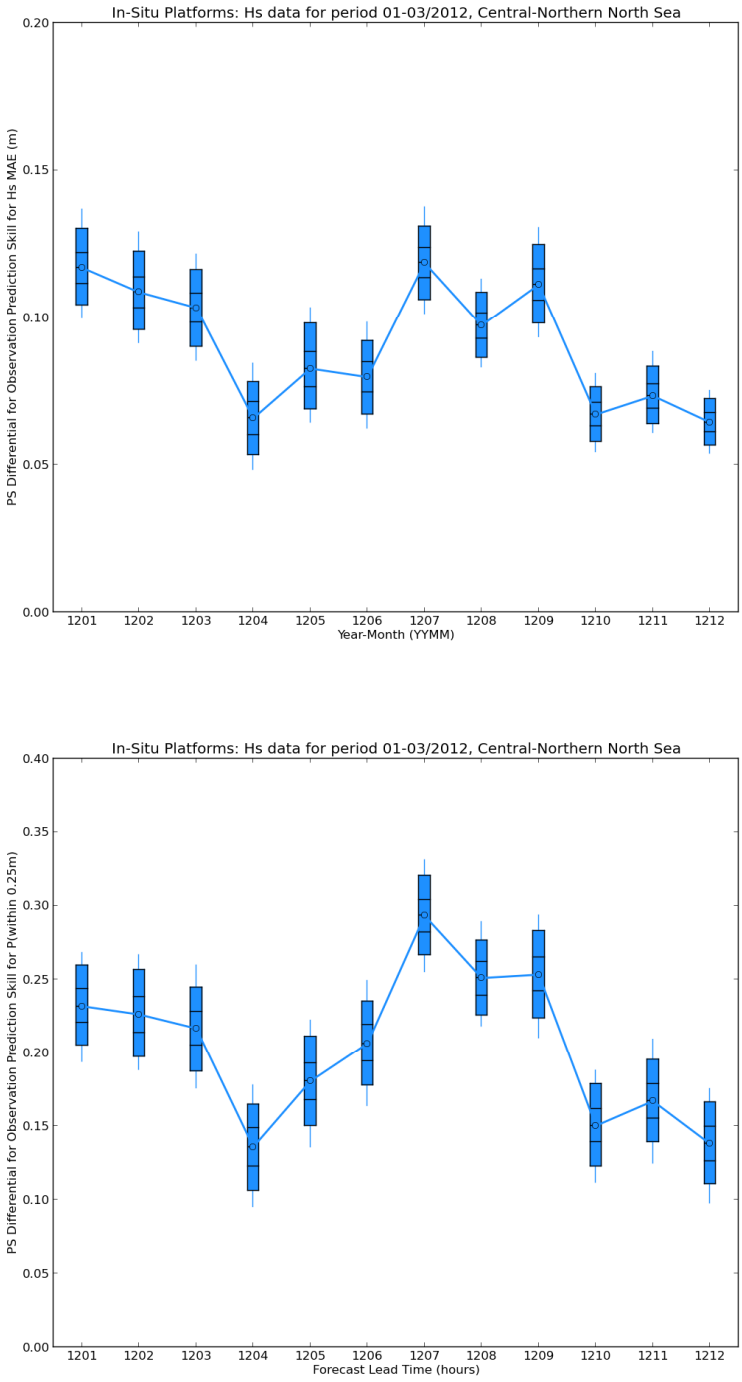


Figure 3.12. Idealised versus direct comparison observation prediction skill differential for (upper panel) MAE and (lower panel) probability of prediction within 0.25m of reference, for rolling 3 month samples during 2012 using model (T+0) against in-situ data.

Figure 3.13 shows a comparison of SNRMSE derived against both in-situ and satellite references. Prediction skill against each reference is shown explicitly on the x- and y-axis and the difference to the idealised case can be seen by judging the distance between the direct comparison symbols and the idealised case symbol. In the plots the green symbol marks the idealised case, and the blue symbols mark direct comparisons at different lead times. Cross-centres mark the ensemble mean value and the cross extents lie at the 5% and 95% quantiles. The plot shows that the prediction skill in the idealised case is similar against both in-situ and Jason-2 references, that the model data conform somewhat more closely with the in-situ data than the Jason-2 data and that skill diminishes significantly with lead time.

Figure 3.14 applies the method to the Success Ratio statistic for 2m wave height exceedence threshold. In this figure the performance of the model (over background chance) against both references is very similar and skill diminishes with increasing lead time. As for the SNRMSE the direct comparison data fall closer to the idealised case for the in-situ data than for the Jason-2 data, although the Jason-2 sample size can be seen to have considerable impact on the range of values that the observation prediction skill takes (extent of the blue crosses).

III.5 Discussion

The method shown in this document presents verification that can be used in a quantitative sense, or as a qualitative method of assessing model skill in performing particular tasks through contextualising the metrics generated from direct comparison between prediction and observation against idealised and naive scenarios. The adoption of this tiered structure to the verification process has been established based on MyWave user preferences to see the direct comparison and to retain a separation between data derived against satellite and in-situ sources. This is a pragmatic choice, but also one where effects from the choice of observation error model used in the idealised case can be viewed explicitly, and the verification is not ‘over-processed’ in trying to compensate for the observation errors. By adopting a simulation (rather than analytical) approach the method should be generically applicable to the majority of metrics proposed in MyWave-D4.2a.

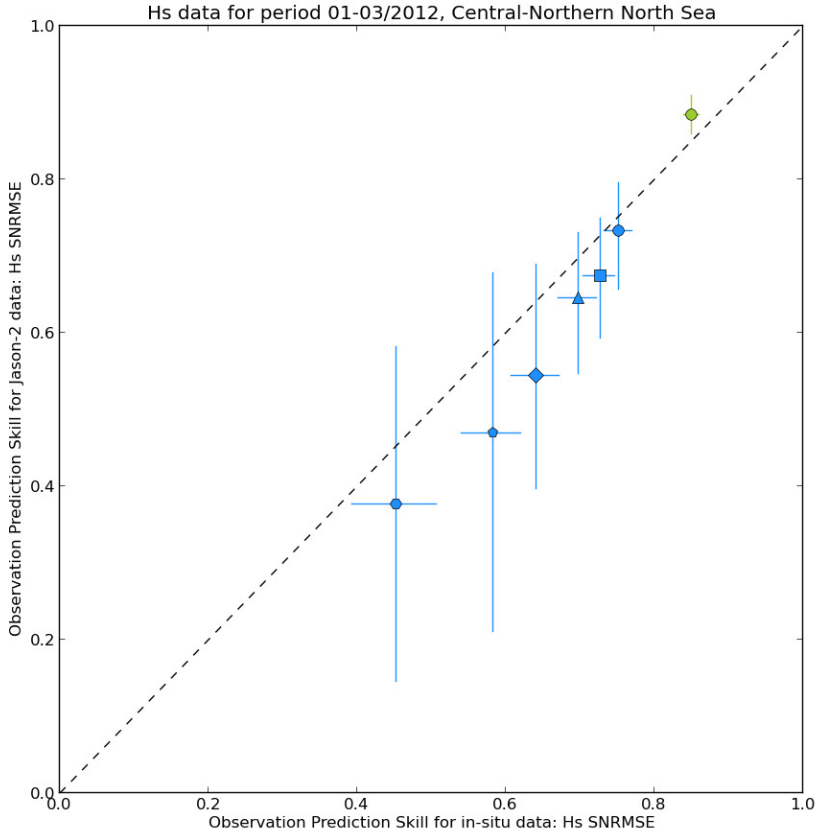


Figure 3.13. Observation prediction skill comparisons for SNRMSE at varying model lead time (blue, 0-circle, 24-square, 48-triangle, 72-diamond, 96-pentagon, 120-hexagon) and idealised prediction (green). X-axis data are comparisons against in-situ observation and y-axis data are comparisons against Jason-2.

Visualization and description of the results shown in subsection III.3 indicate that the contextual data has utility in helping explain the metric values meaning. A skill score approach is necessitated in order to compare verification derived against different backgrounds, but the tiered nature of the verification proposed means that users not interested in this aspect of the data would not have to use it. The next project step is to make further trials of the visualization method with interested user groups.

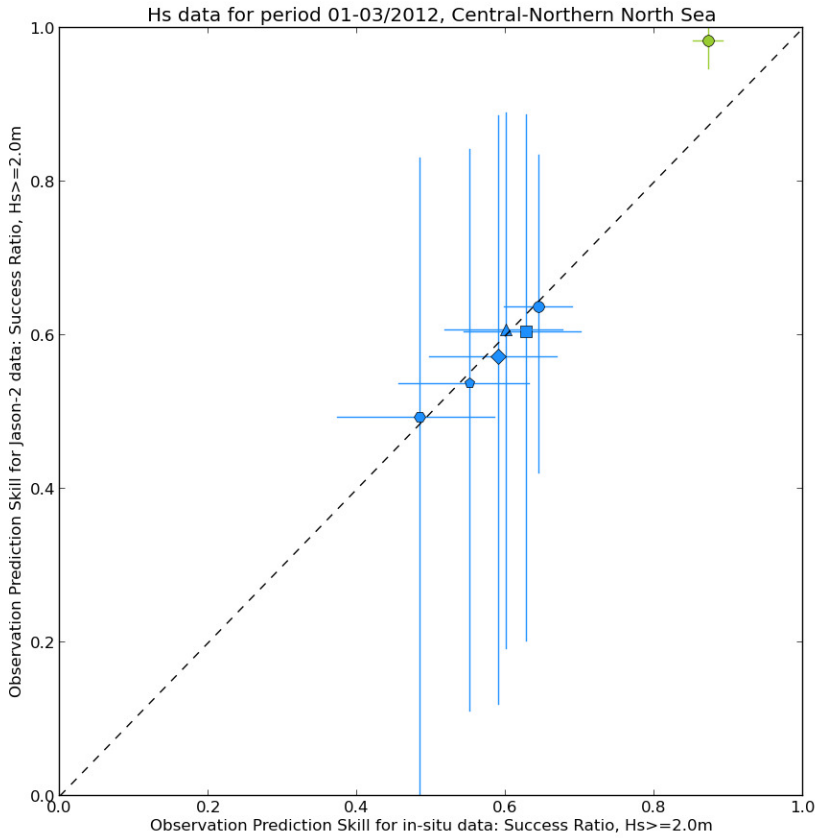


Figure 3.14. Observation prediction skill comparisons for Success Ratio against a 2m threshold at varying model lead time (blue, 0-circle, 24-square, 48-triangle, 72-diamond, 96-pentagon, 120-hexagon) and idealised prediction (green). X-axis data are comparisons against in-situ observation and y-axis data are comparisons against Jason-2 altimetry.

IV SUMMARY AND NEXT STEPS

This report has demonstrated the application of a triple collocation method to establish regional descriptions of observation error characteristics and has proposed a method to incorporate both observation error data and sample variability within verification that uses the observation as reference data. The incorporation of these factors within verification is important as it lends crucial context to verification results, particularly in aiding the user to understand what 'good' performance should look like once errors in the observing systems have been taken into account.

Regional triple collocation assessments of observation errors in two European regional seas, the North European Atlantic Margin (NEAM) and the North Sea, demonstrated that robust and consistent estimates (within +/-1%) of both in-situ and satellite altimeter errors could be generated from comparatively small data samples using the method of Janssen et al. (2007). The main issues found in the study were a susceptibility of the results to outlying data (from poor in-situ observations), changes in the in-situ network and, at the scales analysed, that altimeter observation errors are subjectively controlled by the choice of super-observation used. Nevertheless, relative stable of in-situ and altimeter error estimates were obtained despite the model data used being inhomogeneous. Estimates of (relative) random observing errors for both NEAM and North Sea were consistent within 1-2%, at close to the 10-12% level for the in-situ data and 5-7% level for the satellite data. However, the linear calibration between satellite and in-situ data (as the reference) was more variable between the sea areas (1.02-1.04 in the NEAM and 1.06 in the North Sea), suggesting that some regional variability in the data should be accounted for when citing observation errors. The largest variations were introduced by changes to the in-situ network changed over time, which suggests that if observation errors are to be applied in a quantitative sense in operational verification then regular review of triple collocation results with contemporary observation data needs to be made.

The proposed verification methodology has been strongly influenced by user feedback that stated a preference to see quantitative direct comparisons between forecasts and observations and to retain a distinction between verification against different observation types, but which did express an interest in the addition of information showing the effects of background sample and (to a lesser degree) observation error. This has led to adoption of a tiered approach in which presentation of direct comparisons between the forecast model and

observations are made and the variability of the metric is assessed through re-sampling; a comparative score based on idealised forecast performance, (for which the only errors are associated with the observation) is synthesized to provide a baseline estimate of what good looks like; and a similar naïve comparison is provided to give a lower reference bound. Since the direct comparison and idealised/naïve baselines are generated from the resampled matchup data and a simulation process, this method can be robustly applied to a large number of metrics identified in previous MyWave WP4 reports. Intercomparison of performance against different baselines can be assessed subject to normalising the data in order to account for background sample variability.

Subsequent work within the remainder of the project will undertake a second round of consultation with users as to the effectiveness of the proposed verification method, required steps to refine presentation and provide metadata to help interpretation, and to identify which metrics are best presented in this manner. The application of observation errors defined in the regional triple collocation study will also be extended to wave Ensemble Prediction System verification within Subtask 4.3.1 and tasks within MyWave WP3.

V REFERENCES

Bidlot J.-R., J.-G. Li, P. Wittmann, M. Faucher, H. Chen, J.-M, Lefevre, T. Bruns, D. Greenslade, F. Ardhuin, N. Kohno, S. Park and M. Gomez, 2007: Inter-Comparison of Operational Wave Forecasting Systems. Proc. 10th International Workshopon Wave Hindcasting and Forecasting and Coastal Hazard Symposium, North Shore, Oahu, Hawaii, November 11-16, 2007.

Bowler, N.E., 2006: Explicitly accounting for observation error in categorical verification forecasts. Monthly Weather Review, 134, 1600-1606.

Carlstein, E., 1986: The use of subseries methods for estimating the variance of a general statistic from stationary time-series. Ann. Stat., 14, 1171-1179.

Durrant, T.H., D.J.M. Greenslade and I. Simmonds, 2009: Validation of Jason-1 and Envisat remotely sensed wave heights. J. Atmos. Oc. Tech., 26, 123-134. doi:10.1175/2008JTECHO598.1

Efron, B., and G. Gong, 1983: A leisurely look at the bootstrap, the jacknife, and cross-validation. Am. Stat., 37, 36-48.

GlobWave Wave Data Handbook, 2012.
http://www.globwave.org/content/download/10362/68974/file/GlobWave_D.9_WDH_v1.0.pdf

Janssen, P.A.E.M., S. Abdalla, H. Hersbach and J.R. Bidlot, 2007. Error estimation of buoy, satellite, and model wave height data. J. Atmos. Oc. Tech., 24, 1665-1677. doi:10.1175/JTECH2069.1

Kunsch, H.R., 1989: The jacknife and bootstrap for general stationary observations. Ann. Stat., 17, 1217-1241.

Mentaschi, L., G. Besio, F. Cassola and A. Mazzino, 2013: Problems in RMSE-based wave model validations, Ocean Modelling, Dec 2013, Pages 53-58, ISSN 1463-5003

Palmer, T. and A. Saulter, 2013: Assessment of significant wave height correlation distances in the North Sea and North East Atlantic using a mesoscale wave hindcast. Met Office Technical Report (under review).

Saetra, Ø. and J.-R. Bidlot, 2004: Potential benefits of using probabilistic forecasts for waves and marine winds based on the ECMWF ensemble prediction system. Weather and Forecasting, 19, 673-689.

Saetra, Ø., H. Hersbach, J.-R. Bidlot, D.S. Richardson, 2004: Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability. Mon. Wea. Rev., 132, 1487–1501.

Tolman, H.L., 1998: Effects of observation errors in linear regression and bin-average analyses. Q. J. Meteorol. Soc., 124, 897-917.

Tolman, H.L., 2009: User manual and system documentation of WAVEWATCH III™ version 3.14. NOAA / NWS / NCEP / MMAB Technical Note 276, 194 pp + Appendices.