Project - MyWave

Definition of experiment plan and resources for MyWave Task 4.1: Identify 'compatible metrics' using remote sensed and in-situ wave measurement basleines

Reference: MyWave-D4.1

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:** V1.0
Contributors : Tamzin Palmer, Andy Saulter	
MyWave version scope : All	
Approval Date : 28/09/2012	**Approver:** Andy Saulter
Dissemination level: Project	

DOCUMENT

VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:	Andy Saulter	WP4	28/09/2012
Distribution			
	Oyvind Saetra (MetNo) – Project lead, for onward distribution		
	Jean Bidlot (ECMWF), Marta Gomez (PdE) – Collaborating institutions in WP4		

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By
0.1	21/09/2012	all	First draft of document	Tamzin Palmer	Andy Saulter, Jean Bidlot, Alistair Sellar
1.0	28/09/2012	all	Document finalization	Tamzin Palmer	Andy Saulter

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABREVIATIONS

ASAR	Advanced synthetic aperture radar
FAR	False alarm rate
IOC	International Oceanography Council
Hs	Significant wave height
JCOMM	Joint WMO-IOC technical commission for oceanography and marine meteorology
JEA07	Janssen et al. 2007
MEP	Model event probability
MR	Miss rate
NAEM	North Atlantic European margin
REP	Reference event probability
RMSE	Root mean square error
SI	Scatter index
SR	Success rate
WMO	World Meteorological Organisation

APPLICABLE AND REFERENCE DOCUMENTS

Applicable Documents

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

Reference Documents

	Ref	Title	Date / Issue
DR 1			

I INTRODUCTION

1.1 Task aims and objectives

In order to effectively make decisions on the basis of weather or marine data, users must understand not only how the data they are using relates to their individual operating or warning criteria (for example a forecast wave height exceeds a threshold level), but also be able to place a level of confidence on that data. This 'uncertainty information' should be considered as a critical component of any marine service since it allows decisions to either be taken with high confidence or with mitigating actions in readiness.

MyWave WP4 aims to propose common metrics and reporting methodologies that will, within the framework of a future waves core service, allow both the impact of scientific improvements in wave products to be understood and enable users to quantify uncertainty in wave products and apply this to their specific decisions or downstream information tools. The proposed core service verification system will be expected to fulfil the following criteria: (i) exploits both satellite and in-situ observations as comparative truths for wave model forecasts, (ii) allows consistent presentation of uncertainty information for all regions and products, and (iii), fulfils both scientific and practical needs for uncertainty information.

Specific to this report, Task 4.1 will examine issues associated with the provision of consistent uncertainty information based on verification that uses a mix of both satellite and in-situ observations of the true sea-state. The aim is to describe the 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 the task will be to propose a set of measures that will ensure a Marine Core Service 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 baselines.

As described in the original project scope, Task 4.1 is broken into two subtasks, namely:

- Subtask 4.1.1: Triple collocation methods for verification.
  This subtask will use 'triple collocation' verification methods (in which model data is used to bridge temporal and spatial gaps between remote sensed and in-situ measurements) in order to cross compare performance of one or more wave models against in-situ and remote sensed observations in at least one selected European area where both observation types are available in sufficient volume and where wave process representation should not adversely affect either observation method.

- Subtask 4.1.2: Identification of compatible metrics.
  The subtask will use the triple collocation study results to identify model performance metrics that are compatible for both remote sensed and in-situ measurements in the sense that they can be commonly derived for each data type and allow robust comparison of results within an integrated verification suite using both forms of observed truth. These metrics will be communicated to other work packages by UKMO for use in their verification components.

1.2 Structure of report

The purpose of this report is to: document the background research upon which the triple collocation and metrics assessment work will build; detail the methods being adopted in this study; describe the available model and observation resources that will be used; and provide information on expected study timelines and outputs.

As a result, the report is structured as follows. Section 2 presents an overview of the proposed work for Subtasks 4.1.1 and 4.1.2 and its relation to existing peer reviewed research. Section 3 provides a more detailed description of methods and metrics to be used,

plus the data resources expected to be needed to complete the work. Section 4 summarizes the work plan and details expected timescales for delivering the research and its outputs.

2. STUDY OVERVIEW AND BACKGROUND

2.1 Motivation

Numerical models of the atmosphere and ocean surface waves are a key component in modern day forecasting of sea conditions. Development, validation and (in cases) initialisation of wave models are all based on achieving an optimal agreement between the model and observed estimates of the true sea-state. In a global context the observations are comparatively sparse and the availability of data is further limited by the range of wave characteristics that can be readily and reliably measured, which necessitates that observed data are exploited as effectively as possible. A case in point is operational verification of forecast models which is conducted using finite sampling periods, typically of the order of weeks to a year in length. Ideally such verification is made against a range of conditions relevant to sea areas for which the model will be applied and that this is based on samples that are statistically viable in terms of both data volume and independence.

Operational wave model forecast verification is dominated by use of paired model-observation data, with the observed datasets mainly comprising measurements from in-situ sensors and satellite mounted remote sensing instruments. The predominant parameters measured (in terms of data volume) are wind speed and direction, and significant wave height (Hs) plus various forms of wave period and direction. As will be discussed in the following section, both forms of observation have constraints in terms of their data coverage and in addition may include differing levels of observation error. These errors might, when taken in isolation, bias our view of the overall performance of a forecast. Using the different measurement datasets in combination offers some alleviation of the data coverage issue. When this is done the usual situation is to see results from independent verification studies using the two measurement types cross compared and discussed (e.g. Reistad et al., 2011; Ardhuin et al., 2011). Whilst the approach is entirely appropriate for peer review of model performance in scientific literature, the MyWave project needs to consider that verification results will be presented to downstream users that have less experience in discriminating between such data. Interpretation of the verification should be simple for these users and it is important to avoid the generation of potentially conflicting results derived from datasets

with varying levels of coverage by different observation types in different regions. Assessment of the potential to produce consistent 'single source' verification for defined sea areas within operational sampling periods is therefore required. Such an assessment has two components: understanding and alleviating the relative impacts of differing observation errors on the model verification, and defining if and how operational data sampling in various sea basins might be enhanced by using a combined resource.

2.2 In-situ and satellite remote sensed observations: uses and issues for operational wave model verification

In-situ measurements encompass a wide variety of buoys and platforms with different sensor types. Floating 'buoys' range from relatively small diameter (1-2m) spheres to lightvessels, and sensor types include heave sensors, X-band radars and downward pointing lasers. Whilst measurement types and methods may vary, common properties of the data are a fixed position in space, the measurements are (in principle) continuous in time, and waves are sampled over a 10-30 minute period. The measurements are therefore representative of a sample of individual waves passing a point with the number of waves measured dependent on wave speed, e.g. in deep water a 20 minute burst sample represents a 6km spacing for 6 second wave energy and a 19km spatial coverage for 20 second waves. The main weakness in the in-situ network is geographical data coverage. The majority of buoys and platforms are located with a few hundred miles of coastline and 90% of in-situ stations are sited in the northern hemisphere (Figure 1). The proximity of these sites to the coastline could have the impact that the in-situ dataset becomes dominated by measurements made during early (fetch limited) stages of wave growth and of relatively mature wave systems at the ends of long ocean fetches.

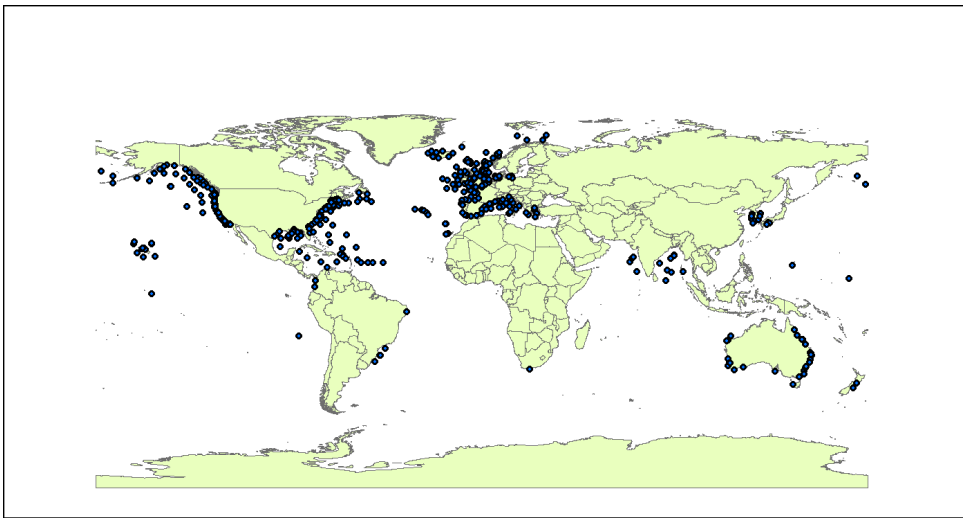

Figure 1. Buoy locations in the JCOMM buoy intercomparison scheme, January 2012.

The period from the early 1990s to present has seen remote sensed measurements from polar orbiting satellites become a major part of the global wave and wind observation dataset. Radar altimeter wind speed and Hs measurement inversion algorithms can be considered mature and the data have been demonstrated to bear a good comparison against in-situ measurements (e.g. Carter et al., 1992; Queffeulou and Croize-Fillon 2009). Advanced Synthetic Aperture Radar (ASAR) measurements can be inverted to generate observations of (truncated) two-dimensional wave spectra, and a number of recent studies and applications have demonstrated the utility and robustness of these data (for a summary of examples see Hasselmann et al., 2012). Over the course of the last 10 years, there have been between 2 and 4 serviceable altimeter instruments in operation and either 1 or 2 operational ASAR (GlobWave wave data handbook, 2012).

Satellite measurements are continuous in space and time as the space vehicle makes its orbit. The sampling characteristics of the instrument are therefore dependent upon the instrument swath, sounding frequency and orbit type. Using the example of the altimeter measurement a sounding at 1Hz is most commonly used, yielding a footprint that covers approximately 6-7km in the along-track direction and with a diameter 2-10km increasing with the sea-state (since the backscatter increases as waves get bigger and wavelengths longer). Mission repeat cycles are dependent on satellite inclination, and vary between 10 and 35 days. The choice of instrument also affects both the maximum latitude at which the

instrument measures and the longitudinal distance between repeated swaths. Thus altimeter soundings are equivalent in nature to the shorter frequency wave energy sampling at a buoy, collectively have greater overall spatial coverage at high resolutions along-track, but provide low temporal coverage at fixed locations (e.g. Figure 2). The extra spatial coverage increases the opportunity for satellite instruments to sample a variety of wave conditions compared to the in-situ network globally, but the lack of temporal coverage means that satellite data needs aggregation over a long period in order to obtain a climatologically representative sample in smaller areas.

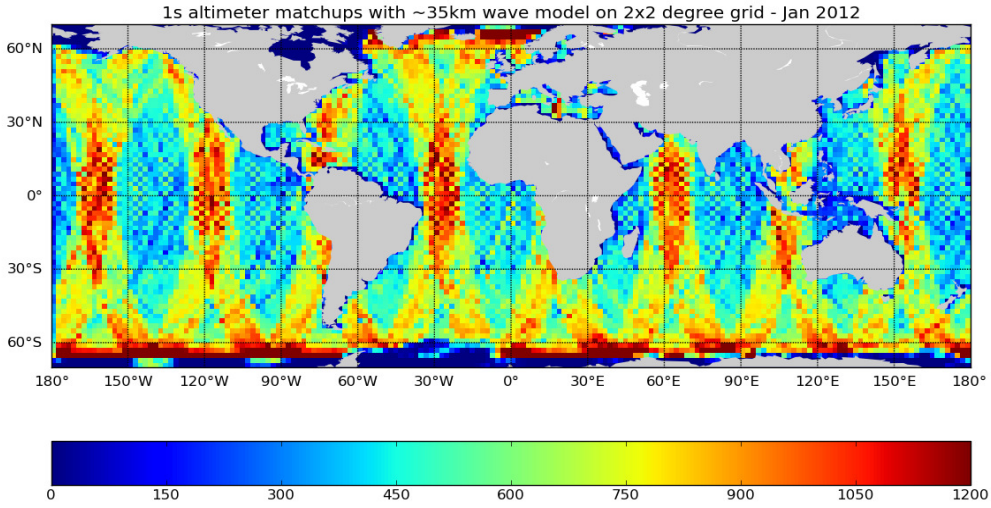

Figure 2. ESA GlobWave project wave forecast verification scheme matchup sample numbers for 1Hz altimeter soundings from Jason-1, Jason-2 and Envisat missions during January 2012. The data were sampled at 1Hz and then aggregated onto a 2x2 degree grid. Assuming individual passes to be independent and comprising approximately 20-25 soundings (based on an 8km along-track resolution per sounding), observations at offshore locations globally were generally made independently between 10 and 60 times in that month.

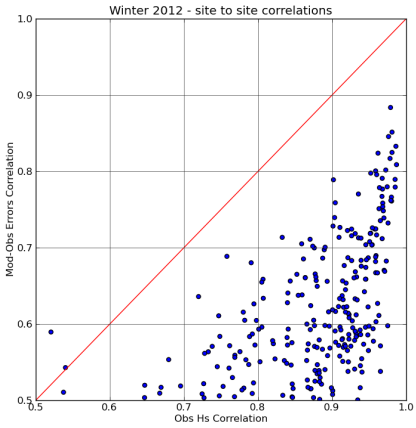

Figure 3. Site to site correlations of (x-axis) observed Hs and (y-axis) model-observation Hs errors for buoys in the JCOMM intercomparison. Data were processed over a 3 month sample.

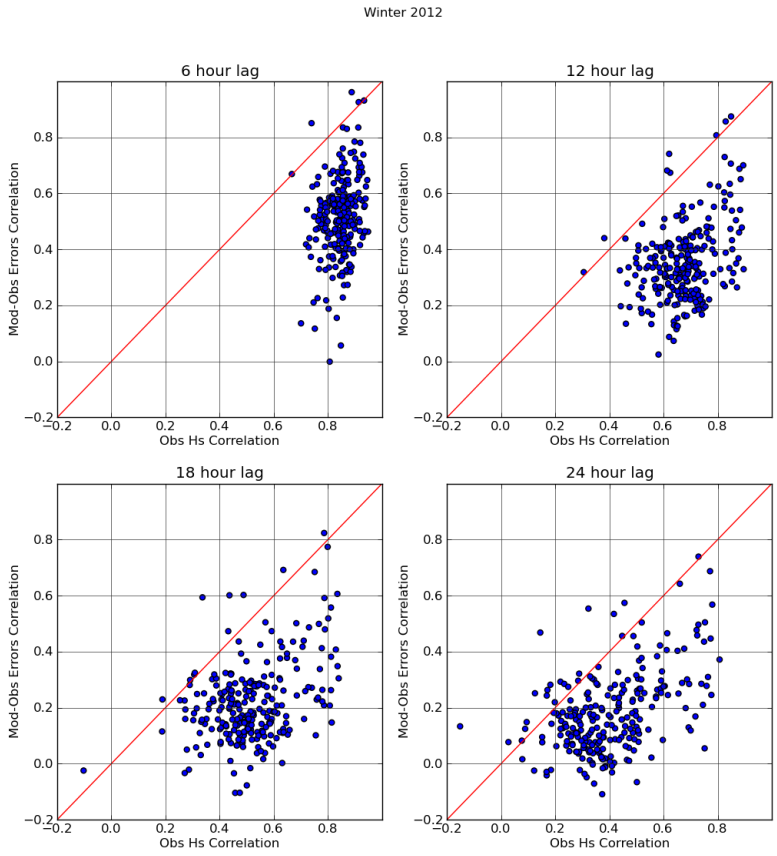

Figure 4. Temporal correlations of (x-axis) observed Hs and (y-axis) model-observation Hs errors for buoys in the JCOMM intercomparison. Data were processed over a 3 month sample.

In order to correctly sample observed data for the task of verification, in which statistics are usually generated based on assumptions regarding data independence, some thinning of the dataset may be required. For example, an assessment of spatial and temporal independence in the in-situ dataset used by the Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology intercomparison of operational ocean wave forecasting systems against buoy data (Bidlot et al, 2002; known elsewhere in this document as the 'JCOMM buoy intercomparison'), suggests that not all sites in the network sample independently (Figure 3) and that a temporal spacing of 12-18 hours between measurements is needed to ensure low correlation (Figure 4). Janssen et al. (2007) cite an along track correlation lengthscale of 70km for the ERS-2 altimeter measurements. When this extra constraint on sampling is applied limitations in coverage of conditions over a finite (weeks to months) sample period can be envisaged for both measurement types.

In summary, whilst enormous efforts are undertaken to provide wave measurement datasets worldwide, the volume and consistency of data coverage from individual networks are not entirely ideal, measurement errors between networks have a high potential to be variable, and the task of verification may only be able to benefit from a subset of the observations. An optimal verification methodology will need to be mindful of these limitations and do its best to mitigate them. Including metrics and methods that are reasonably insensitive to the type of observations used is one desirable property of such a system (Bowler, 2006).

2.3 Observation errors and effects on verification

2.3.1 Sources of error

Model verification using observed data is based on various analyses of 'matched pairs' of the two data types. Both the model and observation provide an estimate of the true conditions, which means that the overall error distribution described by the population of matched pairs will include both model and observed error contributions. It is often assumed that any errors inherent in the estimate of the true conditions by the observations are small, or that the same observed error distribution applies to separate elements of the observing network. In this

latter case the model is assumed to be referenced against a consistent baseline. However a number of recent studies, including Janssen et al. (2007) and Durrant et al. (2009), have described significant errors within the global dataset of in-situ measurements, attributable to the array of measurement and data processing techniques used. Durrant et al. (2009) further noted wave state dependent biases between altimeter and buoy data. The inference from these studies is that neither of the assumptions regarding magnitude and self-consistency in the observed data can be taken for granted and that the observation errors may impact verification as a result.

Differences in errors from in-situ and satellite data, including the different regional characteristics demonstrated by Durrant et al. (2009), also present an issue for any attempts to combine the two data types in order to generate a single matchup sample. In addition the nature of the observed data population cannot be assumed to be held constant with time as different networks and processing techniques are introduced. The ideal position, from the perspectives of consistently improving model performance and providing generically applicable estimates of model error suitable for use downstream, is to make an ongoing account of observation error populations and use these data to help isolate estimates of model errors versus a 'best possible' simulation of the true environmental conditions.

In addition to errors directly associated with the model and observation, a third contribution the matched pair error comes from differences in the representation of sea-state by the model and observations resulting from variations in the temporal and spatial scales being modelled and measured. Janssen et al. (2007) describe representation effects for buoy, satellite and model data, which are further generalised here. In-situ and satellite scales for sampling waves were discussed in Section 2.1. The spatial scale of the forcing winds and wave model sets the scaling for wave growth and limits representation of local effects. This is generally assumed to smooth things to a factor of around 2-3 times the model resolution (Janssen et al., 2007). Therefore for a (25-50km) global scale model the representation scale is approximately 60-100km and for a mesoscale model would be in the region of 20-30km.

In attempting to identify model errors some decisions need to be taken regarding the representation issue. Janssen et al. (2007) sought to mitigate this by scaling their data toward the assumed model scale, i.e. using 'super-observations' of an altimeter averaged over approximately 50km along-track, and averaging 4 hours worth of in-situ data (which has an equivalent scale assuming energy propagation speeds for a common wave peak period at around 8 seconds). A similar procedure has been adopted by the JCOMM buoy intercomparison (Bidlot and Holt, 2006). The approach to representation taken within this study is discussed in Section 2.4.

2.3.2 Simple error models and corrections to verification data

Some inferences about the scale of impacts of errors from different observing systems on verification might be made through direct comparison of the verification results achieved when different observed data are used to test forecasts for the same region and sampling period. The work in Task 4.1 will include deriving these data. A more detailed exploration of these impacts, which should also test the feasibility of providing the desired consistency between verification made using different observation sources, can also be achieved by employing a relatively simple model to describe the errors in the estimate of true conditions found in both models and observations.

As an example consider a comparison between an observation defined as an unbiased estimate of the true sea-state (S_t):

$$S_o = S_t + e_o \quad (1)$$

where e_o represents a random error with mean value zero in the measurements, and a model estimate comprising both linear and (zero-mean) random errors

$$S_m = a.S_t + e_m \quad (2)$$

By assuming that the random errors are uncorrelated both with each other and with the true sea-state signal, a number of sample characteristics relevant to verification metrics can be defined:

$$\mathrm{E}[S_o] = \mathrm{E}[S_t] \,,$$

$$\mathrm{Var}[S_o] = \mathrm{Var}[S_t] + \mathrm{Var}[e_o] \,,$$

$$\mathrm{E}[S_m] = a.\mathrm{E}[S_t] \,, \qquad (3)$$

$$\mathrm{Var}[S_m] = a^2.\mathrm{Var}[S_t] + \mathrm{Var}[e_m] \,,$$

$$\mathrm{Cov}[S_m, S_o] = \mathrm{Cov}[S_m, S_t] = a.\mathrm{Var}[S_t]$$

and these can be re-arranged to determine certain model error metrics versus S_t, e.g.

Bias; $\mathrm{E}[S_m - S_t] = (a-1).\mathrm{E}[S_o] \,,$ (4)

Error Variance;

$$\mathrm{Var}[S_m - S_t] = \mathrm{Var}[S_m] - 2.\mathrm{Cov}[S_m, S_t] + \mathrm{Var}[S_t] = \mathrm{Var}[S_m] + (1-2.a).\mathrm{Cov}[S_m, S_o] \,, \quad (5)$$

Mean Square Error; $MSE = \mathrm{Var}[S_m] + (1-2.a).\mathrm{Cov}[S_m, S_o] + (a-1)^2.\mathrm{E}[S_o]$ (6)

where $a = \dfrac{\mathrm{E}[S_m]}{\mathrm{E}[S_o]}$.

If in addition a form is given to the distribution of errors in e_o, e.g. a normal distribution, further metrics can be corrected relative to a revised estimate of true sea-state. Tolman (1998) uses such an error model to discuss corrections to linear regression coefficient (a in the equations above). In addition the paper assessed effects of bin size and observation errors on bin-average statistics and concluded that observation errors introduced spurious nonlinearities and underestimates of extremes. A correction was proposed using the error probability distribution function for the observation based on an iterative process of estimating model errors via a polynomial fit across the full data range. Bowler (2006) demonstrated that the apparent performance of forecast data improved for categorical verification once observation errors leading to mis-categorization of events were accounted for. The correction discussed uses a deconvolution of an assumed observed error distribution from a probability distribution function for the observations in order to reconstruct a 'true' contingency table. Observed

error distribution estimates have also been applied to verification of ensemble prediction systems, for example Saetra et al. (2004) and Saetra and Bidlot (2004).

These types of error models require a number of assumptions to be adhered to, but it is believed that their adoption may be plausible for suitably large datasets. Previous work using similar forms of error estimator (e.g. Janssen et al., 2007) have reached robust conclusions. One potential issue relates to using a sensible data sample size in order to ensure that spurious correlations between 'noisy' errors and the true signal are kept small. In order to test this, data samples of varying size were drawn from a representative distribution of winter wave heights for the North Sea. For each sample size 1000 simulations were then performed in which normally distributed random noise was added to the data sample and the resulting variance calculated and compared to an exact solution which assumed no correlation between noise and sample. Figure 5 shows the variability in both the mean simulated variance and between simulation members versus the sample population size, for noise set using a standard deviation of 10% and 30% of the representative sample mean. The data shows that once the sample size used is of the order of 2000 points, the estimates have low variability and are within 0.05% of the exact solution. Assuming independence between 12-hourly reports of wave height this would, for example, equate to a 3 month sample from 10-12 independently located buoys.

Since this is not an unreasonable amount of data to expect within an area based verification scheme, it is proposed therefore that exploring the potential of this type of error model to improve consistency in verification results derived from different observations is part of the study. The main piece of information required to be determined will then be the observed data error.

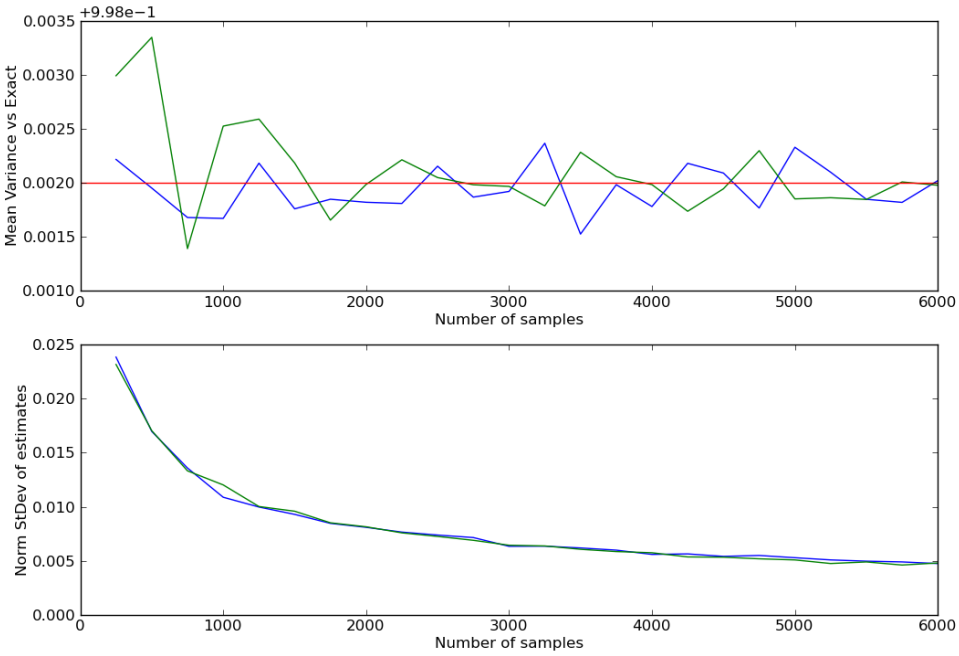

Figure 5. Comparison of simulated estimates of significant wave height (*Hs*) plus (Gaussian) error noise (*e*) sample variance versus an exact solution assuming no correlation between errors and signal (i.e. Var[Hs+e] = Var[Hs] + Var[e]). Top panel shows the mean estimate from 1000 simulations, and the bottom panel shows the standard deviation of estimates – in both cases the data are normalised by the exact solution. Simulations represented by the green line used a error standard deviation set at 30% of the representative mean wave height, and those for the blue line used a factor of 10%.

2.3.3 Triple collocation estimation of observation errors

Central to the application of the error models described in Section 2.3.2 is obtaining an estimate for the distribution of e_o. Methods to derive data describing e_o have been proposed by Challenor and Tokmakian (1999), Freilich and Vanhoff (1999) and Stoffelen (1998). The method of Stoffelen (1998), which derives errors from a dataset of triple collocated independent estimates of an environmental parameter, was adopted by Caires and Sterl (2003) for calibration of wind and wave height data from the ERA-40 re-analysis, and has been further exploited by Janssen et al. (2007) and Abdalla et al. (2011) for error estimates and calibration of fast delivery altimeter data.

A form of the error model described in Section 2.3.2 was used and described by Janssen et al. (2007, hereafter JEA07), who also discuss the assumptions associated with the method. First if the errors in the estimator datasets are assumed uncorrelated, this allows the error of each dataset to be estimated from variance and covariance data. However, making that assumption alone does not allow any calibration of the data and, in order to achieve this, a minimization procedure was adopted which allowed both error and calibration information to be derived following an iterative procedure. JEA07 assumed the form for the estimate of the true sea state to comprise both a linear correction coefficient and a random error part, i.e.

$$S = b.S_t + e \qquad (7)$$

and that b and the variance of e could be reliably estimated from a minimization procedure. In the case of a triple location dataset b is a relative property and needs to be derived using one dataset as a reference, i.e. one estimator is assumed to need no linear correction in order to make an unbiased estimate of the truth. Further details of the method are given in Section 3.1.

In Subtask 4.1.1 (non-assimilative) model data, in-situ measurements and combined mission satellite altimeter measurements will make up the three independent sea-state estimates. Where Subtask 4.1.1 will depart from JEA07 is in the details of the data used. The aim here is to identify error estimates on basin rather than global ocean scales, to use an error representation scale related to the buoy rather than model, apply combined satellite missions, and to use differing model inputs in order to test the effects on the e_o estimates from in-situ and satellite data. In a converse scenario Abdalla et al. (2011) obtained only small differences in model errors when the same model was compared against triple collocations using different satellite data. The choice of basin scales to examine the data may provide some insights into regional variability introduced by measurement/processing methods in different in-situ networks and the prevailing characteristics, e.g. wind-sea or swell dominated, or mixed conditions. The chosen representation scaling is driven by a requirement to focus MyWave verification on downstream usage and is discussed further in Section 2.4.

2.4 Application of observed error estimates to model verification

2.4.1 Parameter choice, sampling and representation issues

The application of the error models will be tested for significant wave height. This choice is based upon the availability of these data as a common parameter from in-situ and satellite observation sources. It should be noted that the aim is not to use the triple collocation component of the work to assess the model errors but simply to quantify errors associated with the observations. Having achieved this the expectation is that the observed error estimates are robust enough to then be used in verification of other models or different data from the model used in the triple colocation (e.g. forecast data).

In order to be based on a sensible triple collocation sample, observed error estimations are likely to need to be derived from relatively long periods (a year or more) in comparison to a standard operational verification period (1-3 months). The application of the observed error estimates to verification data will therefore assume that that the estimates are consistent with the errors contained in the observations during a subsequent shorter operational sampling period. This assumption necessitates testing that the observed error estimates derived from triple collocation are relatively consistent in time, e.g. by comparing observation error data derived over rolling 12 month periods.

The need to consider the verification for MyWave in a user focused manner suggests a need to deal with the issue of representation errors in a different manner to that adopted by Caires and Sterl (2003), Bidlot and Holt (2006) and JEA07. Driven by the need to generate error estimates for data assimilation purposes, JEA07 proposed to mitigate representation errors by super-observing measured data onto a model equivalent scale. In contrast, practical application of operational forecasts will normally be judged directly against spot measurements or often a user's short assessment of sea conditions. Arguably user focused verification should use the same type of method as its reference scale and this study will therefore choose to use an equivalent scale to a 20 minute buoy sample. Satellite altimetry

data will need to be averaged over 3-4 1Hz soundings in order to be similarly scaled. Where the model data to be verified is sourced from a mixture of global, mesoscale or coastal scale configurations, inter-comparisons against such common observed baselines should be a sustainable and universal approach.

2.4.2 Choice of metrics

With the observed error estimated, the next step is to test that, when applied independently to model versus in-situ and model versus satellite comparisons, the observed error estimate and assumed error model will enable consistent description of the model errors. The metrics against which these tests will be made are proposed based on common usage, ability to be corrected using an error model and applicability to the downstream user community. MyWave Task 4.2 will make a review of user metrics, but for considering downstream usage of verification data at this stage in the project it is suggested that a finite demarcation of downstream users into two types, 'forecaster' and 'end-user' can be made. Under these definitions a forecaster will have sufficient knowledge of the model to allow application of verification statistics either in post processing the guidance or communicating risk associated with a given forecast to marine decision makers. The end-user is more interested in verification describing how the model might perform for a specific application and how consistent the model might be with other decision making assets such as on-site observations.

In their simplest form these requirements are covered by the types of 'correctable' metrics discussed in Section 2.3.2. For example bias, standard deviation of errors and root mean square error (RMSE) are common methods to provide an overall measure of systematic and random errors based on the largest possible data sample and provide a useful form of comparison against a climatological or persistence based 'naïve forecast' or between two models. Hanson et al. (2009) introduced an overall performance metric based on the combination of the three statistics in normalised form. Regression relationships and bin-average errors (for which corrections can be derived following Tolman, 1998) can be viewed as a stratification of error data according to the forecast conditions and as such may provide

a more detailed form for comparison of different forecasts and a potentially useful post-processing tool for downstream users wishing to correct model guidance. Categorical statistics, which are generated from contingency tables of correct and incorrect forecasts of a specific event (e.g. Stephenson, 2000) provide a useful indication of performance around specific critical operating or warning thresholds, and lend themselves to application against cost-loss models in order to test hypotheses that the forecast will add value to operational decision making over a long term period. These statistics can be presented in a number of ways with different levels of complexity, but a straightforward collection of measures appropriate to user monitoring of performance, following proposals by the World Meteorological Organisation Offshore Weather Panel, will be adopted for testing in this work. Bowler's (2006) correction method can be applied to these data.

In addition to these metrics two descriptive methods are proposed for examining the data. Quantile-quantile (or q-q) plots are a useful visual method to test the hypothesis that a model is capable of realising the full set of observed environmental conditions without systematic bias, and in this instance also enable a direct comparison between the distribution of data in different observed datasets. The Taylor diagram (Taylor, 2001), based on the combined normalised ratio of standard deviations of two estimators and Pearson correlation coefficient, provides a test of the match between two signals and enables direct comparison between datasets of varying content.

More details of the metrics are given in Section 3.5.

2.5 Testing the potential application of combined measured data

If the use of observation error models is proven to enable a more consistent estimate of the model errors to be made from independently referencing against in-situ and satellite data, the prospect of combined sampling of in-situ and satellite referenced data pairs can be realised. The anticipated benefits of adopting such sampling are increased sample size allowing better stratification of data, providing a unified estimate of model performance, and potentially enabling a resampling by substitution strategy to be introduced.

2.5.1 Sampling and substitution analysis

The simplified example in Fig 6 illustrates how a combined sample might work. The aim is to achieve as large a sample of independent data pairs as possible over the sampling period and area defined. One method to achieve such a sample is to effectively break the area being verified into a set of separate area and time cells and then use data pairs that fall into each of these. The assumption used is that if some cells are more regularly populated than others this will not alias the sample, since all cells sample from the same range of conditions. The validity of the assumption can be controlled by the choice of area and sample period used.

For certain area-time cells the existence of two close in-situ platforms or the collocation of a satellite pass will mean that duplicate model-observation data pairs are available. Where the number of duplicate cell samples is greater than or close to the number of single cell samples the opportunity to resample data by substitution, i.e. generating multiple data samples where the model-observation data pairs are held constant in terms of cell selection but varied by picking randomly from the available duplicates, is presented.

The approach will be tested via the same metrics discussed in Section 2.4. As previously, the main test is that the combined sample dataset can achieve a consistent description of model errors compared to independent in-situ and satellite based assessments. Where the data will differ from stand alone assessments of verification against in-situ or satellite observations is in the necessity to assess a combined distribution of different measurement errors weighted by sample size. In addition a meta-analysis of the combined sample can be made in order to demonstrate whether options for data stratification in the verification are made available.

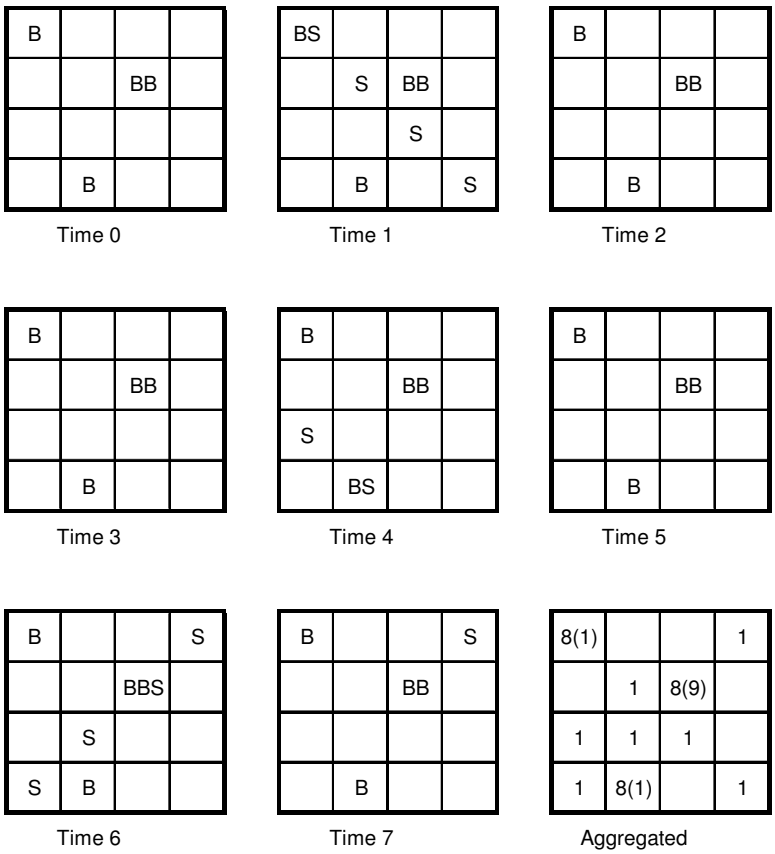

Figure 6. Combined sampling method illustration.

2.6 Study outputs and pull through into MyWave project

From the work discussed in this section Subtasks 4.1.1 and 4.1.2 will expect to output the following:

1. Descriptions of the sampling techniques required to acquire independent samples of raw model versus in-situ and model versus satellite Hs data pairs in European sea areas

2. A comparison of Hs verification statistics generated in the two sea areas using raw model versus in-situ and model versus satellite data pairs over a 'standard' verification period (3 months)
3. Assessments of observation errors in the two sea areas based on a triple collocation approach
4. Results from tests to determine whether the application of established observation error data and a simple error model enables more consistent verification of model errors compared to the raw data in 2
5. A proposed sampling strategy and verification consistency test for paired model-observation data based on a combination of in-situ and satellite measurements.

It is anticipated that these results will enable the MyWave project to:

- Quantify the variability introduced to model verification introduced by using separate observation data sources.
- Determine 'basin scale' errors for in-situ and satellite observations in European waters, and assess their consistency on both a geographical and temporal basis.
- Determine whether it would be valid to include, within a waves marine core service verification programme, statistics that are corrected using the combination of a simple error model and estimates of observation errors. This decision would be based on the ability of such a scheme to generate consistent measures of model errors independent of the observation source.

The success or otherwise of these assessments will enable the project to determine whether verification accounting for the observation errors can be proposed within performance measures for any future waves core service and if such statistics are best presented using independent or combined measurements.

3 STUDY METHODS

3.1 Task list

In order to complete the study as discussed in Section 2 the following tasks will be undertaken:

- Acquire model, in-situ and satellite data to cover the study period. The datasets will need to be of sufficient length to achieve robust samples for triple collocation plus an adjacent verification period, i.e. 2-3 years of data.
- For selected regions (as discussed later in this section) perform correlation lengthscale analyses to determine criteria for sampling of observations and matchup with model data.
- Perform the triple collocation study in selected regions; obtain observation error data and test stability in time.
- Calculate verification metrics for raw model versus in-situ and model versus satellite matched pairs over 'standard' 3 month data periods.
- Test and apply observed error data and the associated error model to generate corrected model versus in-situ and model versus satellite statistics.
- Use a combined data sampling method to generate raw model versus observation verification statistics.
- Using error data and the associated error model, generate corrected model versus combined observation verification statistics.
- Analysis of results and reporting, including proposed method for operational implementation if necessary.

The remaining subsections in this part of the report detail datasets and methods expected to be employed in the study.

3.2 Data for study

In order to use a contemporary period with available resources in terms of model, buoy and satellite data the period from 2009 to present has been selected for this study. The main wave model data that will be used for the triple collocation study comes from a hindcast using the Met Office WAVEWATCH III model. This runs from 2000 to present and has been carried out using an 8km resolution model of the European domain. In-situ data are available in a number of regions within this area, and it is proposed to use hourly data made available to the JCOMM intercomparison project for the triple collocation study. There are three main areas of interest where relatively high densities of in-situ data are available:

- North Sea (3°W - 10°E , 51°N - 63°N)
- Mediterranean (6°W - 36°E, 30°N - 46°N)
- North Atlantic European Margin (20°W- 0°W, 30°N- 65°N)

Fast delivery satellite altimeter data from Envisat, Jason-1 and Jason-2 are available for this period via the GlobWave project. Data from these sources have been used to estimate the number of collocations within the regions of interest. The results of this audit are presented in the following sections. For example Table 1 presents estimated numbers of 6 hourly in-situ measurements and satellite passes for a 3 month period in each of the areas of interest.

In order to ensure that collocation pairs represent independent measurements some subsampling of the data will be required. In particular the proximity of a number of in-situ measurements to each other in the North Sea, means that the individual measuring devices may not provide independent measurements. The proximity of individual satellite measurements in space and time also results in errors between each of the measurements being correlated. For this reason it is the number of satellite passes that can be collocated with an in-situ platform, rather than the number of individual measurements that have been audited for this study. More detailed results using this approach are given in subsequent sections.

Region	Estimated number of in-situ measurements.	Estimated total satellite passes in domain area.
North Sea	19800	407
NAEM	20160	772
Mediterranean	11880	657

Table 1. Estimated numbers of in-situ measurements and satellite passes for a 3 month period in each of the areas of interest.

3.3 Triple collocation

3.3.1 Collocation methods

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 to have results that are statistically representative of the datasets. Errors associated with representativeness are usually addressed by time averaging observations towards the scales represented by the wave models.

A number of examples of global triple collocation studies are available from the literature. In a simple collocation study of wave buoy and satellite altimeter measurements, Durrant et al., (2009) used a collocation criteria of 50km and 30 minutes. This criterion has been regularly used in intercomparison work, for example Monaldo (1998). In a triple collocation study the spatial scale represented by the model wave field also needs to be taken into account both in defining match up criteria and deriving errors. Typically the spatial scale represented by the wave model is assumed to be actually 2 to 3 times the wave model resolution. For example,

JEA07 used a collocation spatial scale of 200km based on an argument that whilst the ECMWF wave model at the time had a spatial resolution of 40 – 55km, smoothing of wind and wave fields introduced in the models reduced variability at short scales such that the model representation scale was about 100km. Working at these scales allowed JEA07 and Abdalla et al. (2011) to use a 2 hour time window for matching in-situ observations to satellite altimeter measurement (and the model values at the satellite location). To achieve similar representation scales in-situ observations were averaged using 5 individual observations taken at a 1 hourly interval. The 2 hour time window for the altimeter observations was within the 4 hour window of the wave buoy observations. Satellite altimeter super observations were constructed from a number of individual operations from a single pass. The time window of the in-situ observations was centred on the time window of the satellite measurements.

An identified difficulty with carrying out triple collocation over larger spatial scales is that it may be the case that the altimeter, model and buoy are not all sensing the same 'truth'. This may occur if there is an island or large change in the bathymetry between the two locations. The issue can be mitigated by comparing the difference between the model outputs at the in-situ platform and the satellite location and discarding the collocation if relative values differ by more than a given percentage (e.g. JEA07; Abdalla et al., 2011). Linear interpolation is also used to reduce collocation errors, for example the model field value may be interpolated towards the location of the wave buoy and altimeter measurement.

For this triple collocation study the aim of establishing regional observation errors, use of an in-situ measurement representative scale, and the high density of the in-situ data in some regions needs to be taken into consideration when selecting appropriate collocation scales. The availability of the data also needs to be assessed to ensure that the criteria chosen will produce sufficient collocations. In the following section a number of different spatial scales have been used to determine both the number of potential collocations and the correlation between the in-situ and satellite measurements for each spatial scale. Different time windows between the in-situ and the satellite measurements have also been used to assess their relative significance.

3.3.2 The JEA07 method for triple collocation estimation of errors

This section provides the details of a method used to determine the respective errors from 3 independent estimates of the truth which has been adopted following JEA07. In order to use this approach assumptions have to be made about the relationship between the model, observations and the truth. A triple collocation essentially provides 3 estimates of the truth, labelled here as X, Y and Z. These will all be referred to as measurements here although in our study one estimate will be made by a wave model. It is assumed that the measurements depend on truth T in a linear fashion, Eq. (8):

$$X = \beta_x T + e_x \ ,$$

$$Y = \beta_y T + e_y \ , \qquad (8)$$

$$Z = \beta_z T + e_z \ .$$

where e_x , e_y, and e_z denote the residual errors in measurements X, Y and Z. β_x , β_y and β_z are the linear calibration constants.

If two types of measurement have a nonlinear relationship to the truth and a linear calibration model is used, errors may be correlated, for example when comparing altimeters that share the same measurement principle. If it is assumed that the linear model is valid and the errors are uncorrelated:

$$\langle e_x e_y \rangle = \langle e_x e_z \rangle = \langle e_y e_z \rangle = 0 \qquad (9)$$

where brackets denote the average over a large sample.

The calibration constants are eliminated by the following new variables:

$$X' = X / \beta_x \ , \ e'_{x'} = e_x / \beta_x \text{, etc}$$

$$X' = T + e_{x'},$$

$$Y' = T + e_{y'}, \qquad (10)$$

$$Z' = T + e_{z'}.$$

Since the primed observations have uncorrelated errors, the truth can then be eliminated to obtain:

$$X' - Y' = e_{x'} - e_{y'},$$

$$X' - Z' = e_{x'} - e_{z'}, \qquad (11)$$

$$Y' - Z' = e_{y'} - e_{z'}.$$

Multiplying the first equation in Eq. (11) above with the second obtains the variance error in X' in terms of the variance of X' and the covariance's of X' and Y', X' and Z'. Multiplying the first and the third equation gives the variance error in Y' and the variance error of Z' is obtained by multiplying the second and third. This gives:

$$\left\langle e_{x'}^2 \right\rangle = \left\langle (X' - Y')(X' - Y') \right\rangle,$$

$$\left\langle e_{y'}^2 \right\rangle = \left\langle (Y' - X')(Y' - Z') \right\rangle, \qquad (12)$$

$$\left\langle e_{z'}^2 \right\rangle = \left\langle (Z' - X')(Z' - Y') \right\rangle.$$

If the errors are uncorrelated this approach can be used to estimate the variance of the error in each of them.

Next a calibration of the measurements can be carried out. The truth is unknown, so only two of the three calibration constants can be obtained. One, say X, is chosen as the reference. The calibration constants for Z and Y can be obtained using neutral regression (Marsden 1999). For example, the regression constant for Y is:

$$\beta_y = (-B + \sqrt{(B^2 - 4AC)}) / 2A \qquad (13)$$

where $A = \gamma\langle XY \rangle$, $\gamma = \langle e_x^2 \rangle / \langle e_y^2 \rangle$, $B = \langle X^2 \rangle - \gamma\langle Y^2 \rangle$, and $C = -\langle XY \rangle$. Y can be replaced with Z in Eq.(13) to give the regression constant for Z.

This calibration will clearly affect the error estimations for X, Y and Z, which will in turn affect the calibration constants and so on. An iterative procedure was adapted by JEA07. Starting with an initial guess of $\beta_y = 1$, $\beta_z = 1$, Y and Z are scaled by the calibration constants; β_y, β_z. Eq. (12) is then used to obtain a first estimate of the errors. The first estimate of the calibration constants is then calculated from Eq. (13) The next step is to scale *Y* and *Z* with the new calibration constants, then determine the errors and regression constants as before using equations (12) – (13), this is continued until the results converge. It is only possible to carry out a relative calibration; however the choice of reference standard will not affect the results (JEA07).

3.4 Independence time- and length-scale assessments

3.4.1 Assessing spatial and temporal scales for collocation

It is important to establish the length scales at which agreement between the in-situ and satellite observations are reduced to the extent that they do not provide a representative measurement of the same sea state. A degree of error due to the collocation method will always be present. Ideally the in-situ and the satellite measurements would be collocated spatially by no more than a few kilometres. However this would result in very few collocations and it would not be possible to create a statistically significant data set. There are a number of methods to enlarge the collocation area. One approach, which was used by JEA07, is to employ the idea of an acceptable collocation error by comparing the difference between the model output at the satellite and the in-situ platform. If the relative error (X_{alt} – $X_{in\text{-}situ}$ / mean wave height) is more than 5% then the collocation error is regarded as

unacceptable as this indicates that the satellite and in-situ measurement are not sensing the same 'truth'.

An assumption that is often made is that for a given radius from a point source, such as a wave buoy, collocation errors will be homogeneous. This is not always the case if conditions are particularly variable within the collocation area (Greenslade and Young 2005). For example whilst in one direction a distance of 100km may be acceptable, large collocation errors may exist with 10's of km in another. This is particularly likely to be the case in coastal areas where local bathymetry and the coastline itself may create rapidly varying sea-states. Islands may also cause wave blocking or focusing in some areas. In a regional collocation study it will be necessary to use observations in areas closer to the coast than might be used globally. The agreement between the in-situ and the satellite data may vary considerably for the same spatial length scale depending on whether it is in deep water in the open ocean or in relatively shallow water close to the coast. A comparison of how different spatial scales may affect both the number of collocations along with the effect on the collocation error has also been carried out.

The analysis determined the correlation between the in-situ and satellite data with different spatial and temporal scales for the North Sea, North Atlantic European Margin (NAEM) and Mediterranean. In the first analysis, matches between the in-situ data and satellite measurements were made where they occur within a 1 hour window of the in-situ record. The in-situ data used were available at 3 hourly intervals. The correlations between the in-situ and satellite measurements were calculated based on satellite observations taken within a 50km, 100km and 200km radius of the in-situ platform, using data from 2009. Initially the average of the first three satellite measurements for each pass within the selected radius and that occurred within 1 hour of the in-situ record was used, the rest were discarded. In the actual study the closest measurements to each in-situ platform will be used rather than the first measurement from a pass to fall within the given radius of that location. Therefore the collocation estimates of correlation made here are likely to be conservative. In a second analysis satellite observations that occurred within a 3 hour window of the in-situ record were used, the analysis method was otherwise identical. Using a more relaxed time window resulted in many more collocations during the sample year; however it did result in a reduction of the correlation between the satellite and in-situ data.

3.4.2 Results of correlation analysis

A comparison of the number of collocations for each method is given in Table 2 and 3. Increasing the time window to 3 hours resulted in almost 3 times as many spatial collocations between the in-situ data and satellite. As the in-situ data used in this audit is 3 hourly, an analysis using hourly records would be likely to result in a significant increase in the number of collocations as seen in Table 2. In addition, the use of 2009 as the audit year coincides with the lowest population of in-situ platforms in the dataset and so the estimates of data volume should be conservative. The limited number of collocations and proximity of in-situ platforms to the coast in the Mediterranean suggested that the study should initially concentrate on the NAEM and North Sea as key regions for assessment.

	Collocations based on radius from buoy (km)		
Region	**50km**	**100km**	**200km**
NAEM	933	1586	2409
North Sea	641	873	966
Mediterranean	130	285	419

Table 2. Number of temporal and spatial collocations for the in-situ platforms and satellite during 2009 based on a 1 hour time window between the in-situ record and satellite measurement.

	Collocations based on radius from buoy (km)		
Region	**50km**	**100km**	**200km**
NAEM	2590	4360	6385
North Sea	2193	2692	2831
Mediterranean	651	1294	1909

Table 3. Number of temporal and spatial collocations for the in-situ platforms and satellite during 2009 based on a 3 hour time window between the in-situ record and satellite measurement.

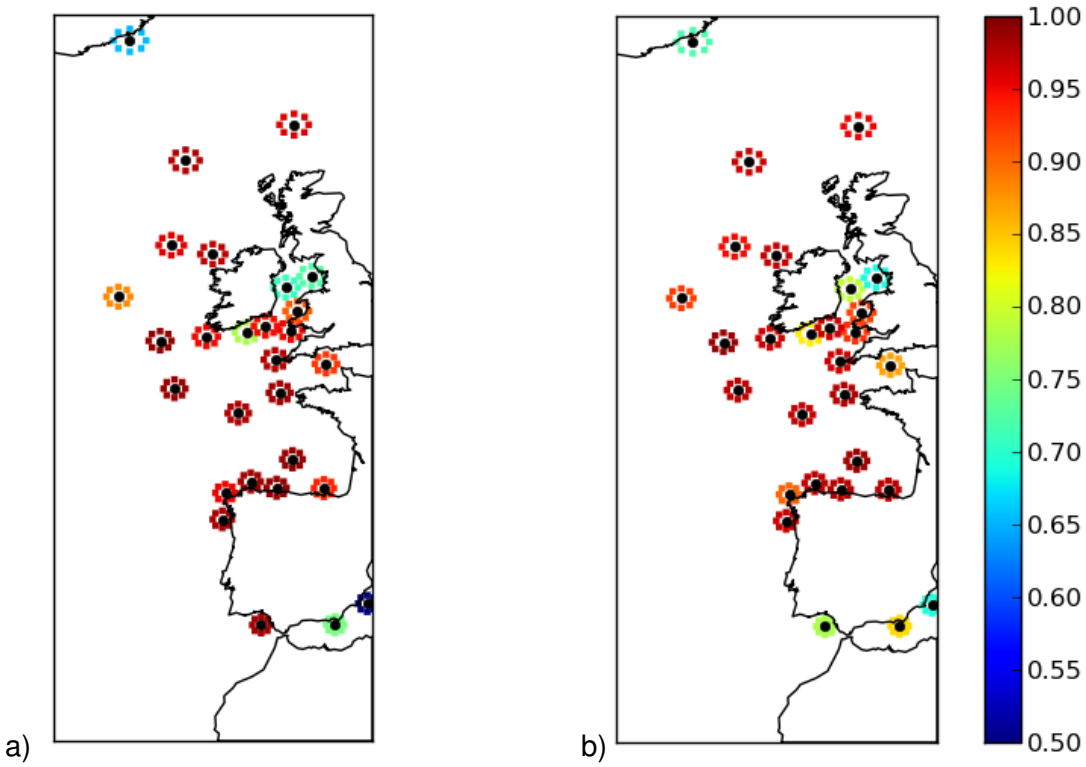

Figure 7. North Atlantic European Margin, correlation (shown on colour bar) between the in-situ platform and satellite for a spatial radius of 50km and temporal window of a) 1 hour and b) 3 hours (the location of the in-situ measurement is given by a black circle).

The correlation between satellite and in-situ data for a spatial radius of 50km in the NAEM is shown in Figure 7. The highest values are seen in the northern areas, west of Ireland and off the coast of Spain. Generally the highest correlation values are seen for areas of deep ocean where conditions are likely to be most homogeneous. The lowest correlations are seen in the Irish Sea, English Channel and the entrance to the Mediterranean. These are the areas where wave conditions may vary on scales of even a few kilometres. Overall an increase is seen where a 1 hourly window has been used, however the correlation for the wave buoy to the west of Ireland is slightly lower. Some of the lowest correlation values are observed in the Irish Sea. This could be due to the relative importance of tidal influences on the wave conditions in this area. Where tidal currents influence wave conditions significant wave height may be expected to vary more rapidly both spatially and temporally and this needs to be considered when selected suitable collocation criteria. Since land contamination of satellite data may be an issue, and bearing in mind that the Irish Sea and English Channel wave conditions are likely to be predominated by short fetch wind-seas as opposed to more regular mixed wind-sea and swell in the remainder of the NAEM region, it is expected that the NAEM study will concentrate on data associated with open ocean waters only.

The correlation between satellite and individual in-situ measurement platforms in the North Sea is shown in Figure 8. Higher correlation values are generally observed in the north. Lower values are in the south and closer to the coast where a larger spatial radius around the in-situ measurement platform would be expected to decrease the agreement between the in-situ data and satellite due to conditions changing more rapidly over a scale of 10s of kilometres. Particular features of this region are large sandbanks shoreward of the in-situ platform locations. It is possible that if the spatial scales are too large then the in-situ platform and the satellite are not sensing the same truth. This is particularly likely to be the case for the platform close to the Thames approaches. The results here suggest that this location may not be suitable for inclusion in a study of this scale. Overall an increase in the correlation is seen of a time window of 1 hour.

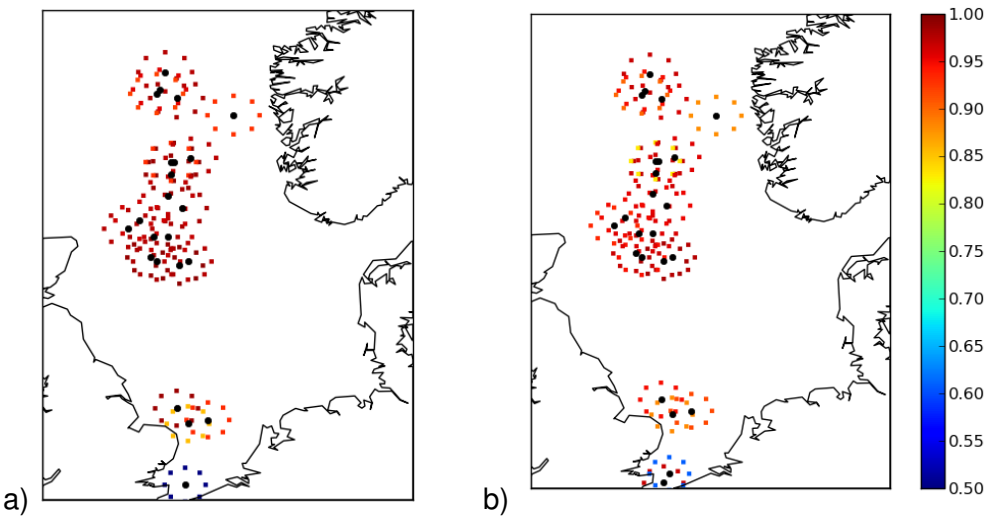

Figure 8. North Sea, correlation (shown on colour bar) between the in-situ platform and satellite for a spatial radius of 50km and temporal window of a) 1hour and b) 3 hours (the location of the in-situ measurement is given by a black circle).

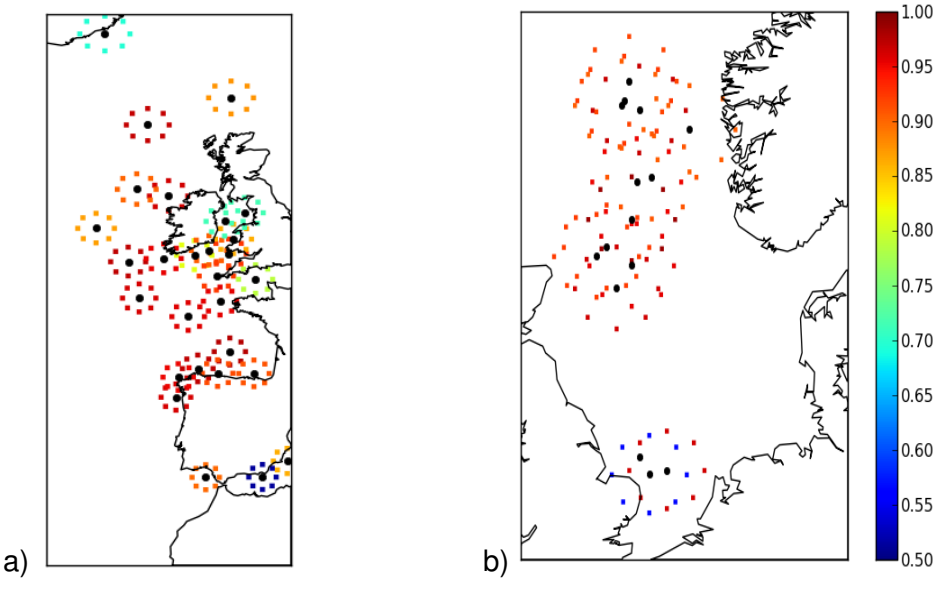

Figure 9. NAEM(a) and North Sea (b) correlation (shown on colour bar) between the in-situ platform and satellite for a spatial radius of 100km and temporal window of 1 hour (the location of the in-situ measurement is given by a black circle).

Figure 9 shows the correlation for a 100km radius at each in-situ platform in the NAEM and the North Sea. There is an overall reduction in the correlation with an increase in the spatial scale in both cases. In particular the reduction is largest in areas such as the west of Ireland (Figure 9a) and to the north and east of Scotland (Figure 9b) where correlation values were previously highest. In the North Sea the density of the in-situ data results in a satellite measurement falling within 100km of more than one in-situ platform on numerous occasions. There is also some overlap if a 50km radius is used, however it is possible to identify separate areas of in-situ and satellite data in this case. In most instances a single satellite pass may be collocated with up to two separate in-situ platforms, but when a 100km radius is used the situation becomes more confused. In all the analyses each satellite pass or super-observation was only collocated with a single platform, in order to avoid duplicate counts being made in the data audit. From the results of this audit a spatial scale of 50km is considered the most suitable for a regional study of the NAEM and the North Sea. From the results of the data audit sufficient data will be available at this spatial scale. A time window of 1 hour produced the highest correlation in the majority of cases. This is the preferred temporal scale.

3.4.3 Calculating covariance fields and estimating correlation length-scales

In order to understand length-scales at which data can be sampled independently over a region, for example to determine methods for sampling satellite observations, a more extensive assessment of correlation throughout the study target regions will be required. To calculate a covariance field between any one location and other points within an ocean basin it is necessary to use model data to provide an estimate as observations are not available over the whole domain. To estimate the degree of correlation between a single location and all other grid cells in a model domain, the covariance field between the grid cell at that location and at all other grid cells in the model domain is calculated.

$$Co = E\left[\left(x - E[x]\right)_i \left(x - E[x]\right)_j\right] \qquad (14)$$

Where $\mathrm{E}[x]$ is the mean and i in this case is the model grid cell at the wave buoy location and j represents all other grid points in the domain.

Covariance data must be generated from a suitable time period, for example one year. For this study 10 years of hindcast analysis data is already available. Once generated covariance or correlation can then be plotted against distance for each buoy location. Functions can be fitted (e.g. Gaussian) to this and used to determine the correlation length scale.

It is unlikely that the correlation length scale will be isotropic at many locations in the regions under consideration in this study. Wave conditions may vary in different directions from an in-situ measurement platform at a point location for a number of reasons. In coastal areas there may be changes in the bathymetry or obstructions due to the coastline or islands. A wave buoy may be exposed to a long fetch from one direction and only short fetch conditions in another. Where a wave buoy is exposed to swell, the correlation length scale is usually longer in this direction (Greenslade and Young 2005). In order to identify where larger collocation errors may occur between the satellite and in-situ measurements it may be helpful to look at variability of the correlation with direction, for example binning results in distance and 10 degree directional bins, similar to method used by Greenslade and Young (2005).

3.5 Metrics

The motivation underpinning the selection of metrics has been provided in section 2.4. A more detailed description of the metrics to be tested and their purpose are provided in this section. Where necessary the metrics will be normalised in order to aid a direct comparison between results drawn from different observed samples in order to test consistency. In addition to metrics used to test consistency of any error correction method the following metadata will also be presented in the study:

Meta.1: Map of area verified (including map of observed sample locations/density)

Meta.2: Model-observation data pair sample size

Meta.3: Sample mean and standard deviation of model (S_m)

Meta.4: Sample mean and standard deviation of reference observation(s) (S_r)

Meta.5: Quantile-quantile (q-q) plot of model-reference observation distributions

The aim of these data are to show a summary of the conditions being verified against, demonstrate that the sample used is likely to be statistically resistant, and provide an assessment of the model's ability to reproduce the observed reference climate.

The following metrics will be used to define general model performance based on variability of (time correlated) errors between model and reference over the full sample range. Normalisations are referenced against the model since, in a predictive system, model estimates of wave conditions are known and the verification data can then be easily applied to understand risk associated with the prediction.

Perf.1: Normalised bias; $NBias = \frac{\mathrm{E}[S_m - S_r]}{\mathrm{E}[S_m]}$ (15)

Perf.2: Normalised root mean square error; $NRMSE = \frac{\sqrt{\mathrm{E}[(S_m - S_r)^2]}}{\mathrm{E}[S_m]}$ (16)

Perf.3: Scatter index; $SI = \frac{\sqrt{\mathrm{Var}[S_m - S_r]}}{\mathrm{E}[S_m]}$ (17)

Perf.4: Taylor plot, which shows departure from perfect replication of (unbiased) reference signal by the model based on position of data using $\frac{\sqrt{\mathrm{Var}[S_m]}}{\sqrt{\mathrm{Var}[S_r]}}$ and Pearson correlation coefficient $\mathrm{Corr}[S_m, S_r] = \frac{\mathrm{Cov}[S_m, S_r]}{\mathrm{E}[S_m]\mathrm{E}[S_r]}$. (18)

Metrics that indicate (time correlated) error levels (and hence options for correction in post processing) are defined below. In all cases the model will be used as the determinant (x-axis) parameter since this form directly provides a functional relationship that can be applied to the model.

Cond.1: Linear slope

Cond.2: Bin-average values of *NBias*

Cond.3: Bin average values of *NRMSE*.

The bin-average statistics offer a different test of errors through the data range without needing to assume a linear relationship. Generally the technique is based on sampling within set variable bins (e.g. every 0.5m of Hs), however this can lead to variability in the sample between bins and potentially misleading results at the extremes of the distribution. Within this study the bin average data will also be tested using an invariant sample size (e.g. overlapping 10% data sub-samples).

Categorical statistics will be generated based on event categories for wave height exceedence of 1m, 2m, 4m and 6m. Tests of forecast success or failure compared to event occurrence or non-occurrence form the contingency table shown as Table 4.

	Event Detected	Event Not Detected
Event Forecast	*a*	*b*
No Event Forecast	*c*	*d*

Table 4 Contingency table for deterministic event forecasts.

From these data various scores can be derived that describe model skill for certain aspects of the forecast. Following requirements for users discussed by the Offshore Weather Panel the proposal is to concentrate on a subset of the most easily understood metrics. It is worth pointing out that in isolation each metric is limited and may reward poor forecasts, so the data should be viewed as a combined set of measures.

Cat1: Model event probability; $MEP = \frac{[a+b]}{[c+d]}$

Cat2: Reference event probability; $REP = \frac{[a+c]}{[b+d]}$

Cat3: Success ratio, the probability of a successful prediction of an event if an event forecast is issued; $SR = \frac{a}{[a+b]}$; note that the false alarm rate $FAR = 1 - SR$ and that the overall probabilities of forecasting then observing an event and raising a false alarm are respectively *SR*MEP* and *FAR*MEP*

Cat4: Miss rate, the probability of not forecasting an event conditional on the event occurring; $MR = \frac{c}{[a+c]}$; note that the overall probability of a missed event is $MR * REP$

Cat5: Odds ratio chance of a correct prediction of either event or non-event status;

3.6 Summary

Section 3 has outlined the triple collocation method that will be used; an audit of the data has been carried out to establish if at least one European area has sufficient data for such a study to be conducted on a regional scale. The results of this audit show that sufficient collocations between in-situ and satellite altimeter measurements occur in at least two areas. These have been defined here as the North Atlantic European Margin (NAEM) and the North Sea. These two areas have similar numbers of wave buoys, 56 and 55 respectively, however the results of the audit show that the NAEM has a larger number of collocations with the satellite measurements. Many of the wave buoys in the North Sea and some in the NAEM are located within 50km of each other, therefore they will not all provide independent measurements. A method of establishing correlation length-scales for these areas was given in section 3.4.3; this will allow independent sampling areas to be defined. Removing measurements that are not independent will lead to a reduction in the number of collocations that are suitable for use in any one study. However areas of higher wave buoy density, such as the North Sea, may prove useful for applying resampling and substitution techniques as described in section 2.4.

4 SUMMARY AND STUDY PLAN

4.1 Document summary

The tasks outlined in this report aim to improve understanding of uncertainty in sea-state forecasts. It is believed this can be achieved by optimal exploitation of in-situ and satellite data, through ensuring consistency of verification statistics derived against either observation baseline and potentially adopting an approach to verification that combines statistics derived against both baselines. As described in section 2 there are a number of different observational systems that sample aspects of the sea-state on different spatial scales. For this reason it cannot necessarily be assumed that different methods of sea-state measurement will provide consistent estimates of wave model errors. This report has outlined a methodology to test whether consistent estimates of model error can be established using different observation baselines, and tested its viability in relation to the analysis period, sampling schemes and availability of data needed to complete the study.

Triple collocation in one or more European areas will be used to estimate errors in the observations and these data will then be used in tandem with their associated error model to establish whether a combination of in-situ and satellite data are able to provide a consistent estimate of model errors. A methodology for a triple collocation study based on that proposed by Stoffelen (1998) and following Janssen et al. (2007) was outlined in Section 3, and it was also established that sufficient observational data was available to make a regional study feasible in at least one European area (Section 4).

The next step will be to investigate in detail the impacts of observational errors on the consistency of a verification scheme that uses a combination of in-situ and satellite data. The use of the same simple error model that underpinned Janssen et al.'s (2007) study to estimate true sea-state conditions for use in the verification process will be explored. The metrics that have been proposed for this study have been chosen for their potential to be corrected using such an error model. As noted earlier the aim of the triple collocation study is not to investigate the model error, but to provide observed error estimators that are robust enough to be used in the verification of one or more models over a given area.

Assuming that the use of observational error models allows a consistent estimate of the model errors to be made independently using satellite and in-situ data, the next step will be to carry out a sampling and substitution analysis based on the combined observation dataset. The resampled observational datasets produced by this analysis will then be compared to models using the same metrics as for independent baseline tests in order to assess the robustness of a combined verification scheme.

Further investigation may also include understanding whether the combination of in-situ and satellite data sufficiently enhances operational verification datasets so as to enable increased stratification of the data. This would allow metrics to be determined for specific conditions such as swell dominated or fetch limited sea-states. Since different study areas may be dominated by different wave conditions a comparison across areas may allow for some assessment of the model and observational errors in different sea-states.

4.2 Study timescales

The intended date for final delivery of work under WP4.1 is December 2013, with reported findings feeding into deliverable D4.3. Following the breakdown of tasks described in Section 3.1, a timetable for the time-scales expected for each of the tasks detailed in this report in Table 5.

Task	OND-12	JFM-13	AMJ-13	JAS-13	OND-13
Data acquisition					
Correlation length-scale analyses					
Raw verification (independent baselines)					
Triple collocation					
Corrected verification (independent baselines)					
Corrected verification (combined baselines)					
Reporting					

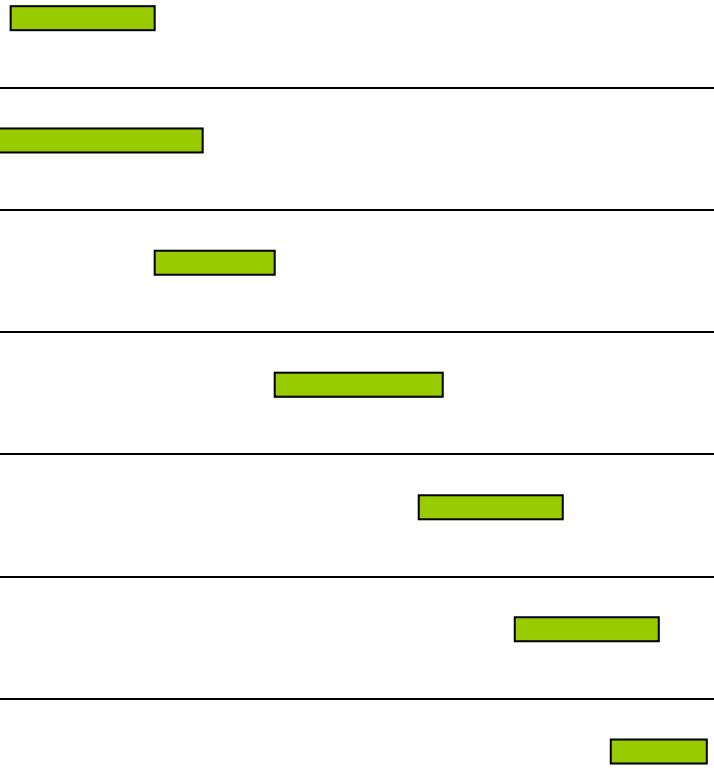

Table 5. Study Plan. OND: October-November-December; JFM: January-February-March; AMJ: April-May-June; JAS: July-August-September

5. REFERENCES

Abdalla, S., Janssen, P.A.E.M. and Bidlot, J.R., 2011. Altimeter near real time wind and wave products: random error estimation. Marine Geodesy, 34, 393-406. doi:/10.1080/01490419.2011.585113

Ardhuin, F., Hanafin, J., Quilfen, Y., Chapron, B., Queffeulou, P. and Orbreski, M., 2011c. Calibration of the 'IOWAGA' global wave hindcast (1991-2011) using ECMWF and CFSR winds. Proc. 12th International Workshop on Wave Hindcasting and Forecasting. http://www.waveworkshop.org/12thWaves/index.htm

Bidlot, J.R., Holmes, D.J., Wittmann, P.A., Lalbeharry, R. and Chen, H.S., 2002. Intercomparison of the performance of operational ocean wave forecasting systems with buoy data. Weather and Forecasting, 17, 287-310.

Bidlot, J.R. and Holt, M., 2006. Verification of operational global and regional wave forecasting systems against measurements for moored buoys. JCOMM Technical Report No. 30. ftp://ftp.wmo.int/Documents/PublicWeb/amp/mmop/documents/JCOMM-TR/J-TR-30/J-TR-30.pdf

Bowler, N.E., 2006. Explicitly Accounting for Observation Error in Categorical Verification Forecasts. Monthly Weather Review, 134, 1600-1606.

Carter, D., Challenor, P., Srokosz, M., 1992. An assesment of Geosat wave height and wind speed measurements. J. Geophys. Res., 97, 11383–11392.

Challenor, P.G. and Tokmakian, R., 1999. On the joint estimation of model and satellite sea surface height anomaly errors. Ocean Modelling, 1, 39-52.

Caries, S. and Sterl, A. 2003. Validation of ocean wind and wave data using triple collocation. J. Geophys. Res., 108 (C3), 3098. doi10.1029/2002JC001492.

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

Freilich, M. H., and B. A. Vanhoff, 1999: QuikScat vector wind accuracy: Initial esimates. Proc. QuikScat Cal/Val Early Science Meeting, Pasadena, CA, Jet Propulsion Laboratory

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

Greenslade, J.M. and Young, I.R., 2005. Forecast Divergences of a Global Wave Model. Monthly Weather Review., 133, 2148-2162.

Hanson, J.L., Tracy, B.A., Tolman, H.L. and Scott, R.D., 2009. Pacific hindcast performance of three numerical models. J. Atmos. Oc. Tech., 26, 1614-1633. doi:10.1175/2009JTECHO650.1

Hasselmann, K., B. Chapron, L. Aouf, F. Ardhuin, F. Collard, G. Engen, S. Hasselmann, P. Heimbach, P. Janssen, H. Johnsen, H. Krogstad, S. Lehner, J-G. Li, X-M. Li, W. Rosenthal, J. Schulz-Stellenfleth, 2012: The ERS SAR Wave Mode – a breakthrough in global ocean wave observations. ESA Publication (in press)

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

Marsden, R.F., 1999. A Proposal for a Neutral Regression. J. Atmos. Oc. Tech., 16, 876-883.

Monaldo, F., 1988: Expected differences between buoy and radar altimeter estimates of wind speed and significant wave height and their implications on buoy altimeter comparisons. J. Geophys. Res., 93, 2285–2302.

Queffeulou P & Croizé-Fillon D June 2009. Global altimeter SWH data set. IFREMER (pierre.queffeulou@ifremer.fr)

Reistad, M., Breivik, Ø., Haakenstad, H., Aarnes, O., Furevik, B., Bidlot, J.R., 2011. A high-resolution hindcast of wind and waves for the North Sea, the Norwegian Sea and the Barents sea . J. Geophys. Res. 116, C05019.

Saetra, O. and Bidlot, J.R., 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, O., Hersbach, H., Bidlot, J.R., and Richardson, D.S., 2004. Effects of observation errors on statistics for ensemble spread and reliability. Monthly Weather Review, 132, 1487-1501.

Stephenson, D.B. 2000. Use of the “Odds Ratio” for Diagnosing Forecast Skill. Weather and Forecasting., 15, 221-232.

Stoffelen, A., 1998. Error modelling and calibration; towards the true surface wind speed. J. Geophys. Res., 103 (C4), 7755-7766.

Taylor, K.E. 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res, 106 (D7), 7183-7192. doi:10.1029/2000JD900719

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., Banner, M.L. and Kaihatu, J.M., 2011. The NOPP Operational Wave Model Improvement Project. Proc. 12th International Workshop on Wave Hindcasting and Forecasting. http://www.waveworkshop.org/12thWaves/index.htm

WMO Handbook of Offshore Weather Services. WMO/TD-No850.