MyWave

Proposal of metrics for user focused verification of deterministic wave prediction systems

Reference: MyWave-D4.2a

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:** 1.1
Contributors : Andy Saulter, Jean Bidlot, Chris Bunney, Marta Gomez-Lahoz, Tamzin Palmer, Paolo Puzzetto	
MyWave version scope : All	
Approval Date : 02 Oct 2013	**Approver:** Andy Saulter
Dissemination level: Project	

DOCUMENT

VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:	Andy Saulter	WP4	02 Oct 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	10 Sep 13	all	First draft of document	Andy Saulter	Jean Bidlot;
					Marta Gomez-Lahoz;
					Tamzin Palmer

1.0	27 Sep 13	all	Document finalization	Andy Saulter	Tamzin Palmer

1.1	01 Oct 13	all	Document errata corrected	Andy Saulter	Andy Saulter;
					Chris Bunney;
					Paolo Puzzetto

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABREVIATIONS

Baseline prediction	Prediction system used as a source of verification comparison
DET	Detection Error Trade-off (plot)
DNV	Det Norske Veritas
Hs	Significant wave height
MAE	Mean Absolute Error
MCS	Marine Core Service
MDir	Mean wave direction
Parameter space metrics	Verification measures that describe characteristics of prediction uncertainty in units of the parameter predicted
pdf	Probability distribution function
Probability space metrics	Verification measures that describe characteristics of prediction performance in terms of probability of a successful or unsuccessful prediction of an event
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
ROC	Relative Operating Characteristic (plot)
SAR	Synthetic Aperture Radar
SI	Scatter Index
SNRMSE	Symmetrically Normalised Root Mean Squared Error
T	(Generic) Wave period
Tp	Peak period of waves
Tz	Mean zero-upcrossing period of waves
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-D4.2b	MyWave: Proposal of metrics for developer and user focused verification of wave ensemble prediction systems	October 2013 / v1.0
DR 2	MyWave-WP4(UC)	MyWave: Categorisation scheme for MyWave users	March 2013 / v1.0

I INTRODUCTION

Tasks in MyWave WP4 will define operational verification methods that can be robustly applied within a wave element of a Marine Core Service (MCS). The purpose of this document is to propose metrics that provide 'user focused' verification data for deterministic forecasts, i.e. information that describes model performance or uncertainty associated with forecasts in an accessible manner. The document also describes a process of user consultation that will be used to ratify the metrics and presents initial findings from this process.

Metrics are identified in two ways. The **purpose** of the metric, i.e. the information that will be portrayed to the user, will be defined in each case and the metrics will also be discussed relative to the **reference data** required. Due to limitations in available observations a truly reliable analysis for waves is not available (Lefevre and Aouf, 2012), so the scope of this document is limited to reference data comprising observations only. Metrics are categorised with respect to reference data as 'common' where data from in-situ and remote sensed sources can be used interchangeably, 'in-situ based' where additional features of in-situ datasets can be applied, and 'satellite based' where additional features of satellite remote sensed data can be applied.

The remainder of the document is set out as follows: in Section II guiding principles that will influence the approach to metrics for MCS verification are discussed; Section III presents the user consultation process that will be used to refine the proposed metrics during the remainder of the MyWave project; section IV proposes metrics be evaluated through the consultation process, these are also summarised in tables presented in Annex A. The metrics which can be made available in an operational system are constrained by data availability and statistical constraints and these are discussed in Annex B.

Outside the scope of this document are a discussion of ensemble prediction system metrics and methods for application of observation error data, which will be subjects of an accompanying MyWave project report in this deliverable (MyWave-D4.2b) and as part of deliverable D4.3.

II PRINCIPLES FOR MCS VERIFICATION

This section defines a set of guiding principles that the metrics and system used for portraying operational verification of wave models should aspire to meet.

Regarding metrics, in order to cover the needs of as wide a range of users as possible, it is suggested that a MCS verification system should comprise metrics which:

- Provide data to indicate whether the prediction system makes a realistic simulation of the observed environment.
- Provide measures that enable users to quantify prediction uncertainty (for the remainder of this document these will be termed 'parameter space metrics').
- Provide information that estimates the probability of success or failure of predictions when MCS model data are used in their raw form (for the remainder of this document these will be termed 'probability space metrics').
- Provide (or enable) comparison of prediction performance against other 'baseline prediction' methods, e.g. a naïve predictor such as random chance or a more sophisticated system such as a new wave model.
- Are considerate of the fact that service users may wish to apply verification data within downstream services or decision making processes.
- Are regularly updated to reflect recent system performance. For example in the MyOcean service metrics are updated every 3 months and are presented in a rolling archive of up to 1 year of data (Alistair Sellar, *pers. comm.*).

In addition MCS verification should consider presentation of the data provided as a critical element of the system. The following requirements are considered desirable:

- Operational MCS verification should provide rapid discovery of metrics that allow downstream users to easily understand performance of the prediction system relevant to their particular use of the MCS data.
- Metrics should be accessible to lay-users; for example, if the metrics provided cannot be explained with a few sentences of text, they are probably not fit for purpose.
- Metrics comparing prediction system performance against a baseline prediction should be meaningful in terms of user decision making.

Adopting these principles means that there is a need to clearly associate given metrics with an application that users can recognise, also to ensure that verification data which can practicably be made available within an operational verification scheme covers as many key user applications as possible. Section IV of this document describes a proposed basket of metrics in these terms.

The requirements also drive or are constrained by a number of technical considerations for verification, which are discussed further in Annex B of this report.

III USER CONSULTATION

III.1 Overview of the consultation process

The MyWave project aims to incorporate user feedback into its final definition of operational metrics and proposal for an MCS verification system (project deliverable D4.4). The approach adopted for obtaining this feedback comprises 3 stages:

Stage 1: Preliminary survey of potential users in order to establish user types and interest in verification information.

Stage2: Detailed survey of verification requirements for users identified as having an interest in verification.

Stage 3: Review of specific metrics and forms for presentation with users identified as having an interest in specific applications of verification data.

The final outcome from this process is expected to be a set of metrics and associated metadata that can be linked to particular user types and have undergone a period of trial and review.

III.2 Initial findings

III.2.1 Stage 1

At writing the preliminary MyWave survey[1] has been provided to 68 potential service users to assess their initial reaction to the project and the concept of a wave component of a Marine Core Service. Responses have been received from 35 users. Questions were included that aimed to identify users based on a hypothetical user categorisation presented in MyWave-WP4(UC). From the responses to these questions an 'in practise' breakdown of users comprises:

[1] http://www.surveygizmo.com/s3/1299480/MyWave-Preliminary-Survey

- *All Scales Developer-Forecasters*: 7 respondents said they worked with wave information from data generation at both global/large regional scales and coastal scales through to provision of forecasts, and that their data and products were used both for planning and operational purposes. These users were split 70%-30% between commercial and government institutions.

- *Coastal Developer-Forecasters*: 9 respondents said they worked with wave information from data generation at coastal scales through to provision of forecasts, and that their data and products were used both for planning and operational purposes. These users were split approximately 60%-40% between commercial and government institutions.

- *Forecasters:* 11 respondents said they worked specifically on providing forecasts and decision aids and, across the group, undertook an even split of tasks focused on marine operations, hazard forecasting and long term planning (using past climatology). These users were split approximately 50%-50% between commercial and government institutions, with one member of the general public also falling into this category.

- *Decision Makers:* 4 respondents said they generally acted as decision makers and, across the group, undertook an even split of tasks focused on marine operations, hazard forecasting and long term planning (using past climatology). These users were split approximately 50%-50% between commercial and government institutions.

- *Developer-Planners*: 4 respondents were involved in niche model development activities at various scales for planning purposes. These users were split 75%-25% between academic and government institutions.

Of these users 25 expressed an interest in further contact on the subject of MCS verification and were split as 6 All Scales Developer-Forecasters, 7 Coastal Developer-Forecasters, 7 Forecasters, 2 Decision Makers and 3 Developer-Planners.

III.2.2 Stage 2

A survey containing more detailed questions regarding user requirements for wave verification[2] was issued on 9th September 2013. Key findings from initial responses (14 users, split as 6 All Scales Developer-Forecasters, 3 Coastal Developer-Forecasters, 3 Forecasters, 1 Decision Maker and 1 Developer-Planner) are that:

- The main requirements for verification data relate to review and intercomparison tasks rather than use in downstream intervention strategies.
- A majority of users would be interested in near-real time monitoring data and downloadable match-up information in addition to review statistics.
- Interactive webpages were considered the best method to deliver verification data.
- Overall wave height, period and direction were considered the most important parameters to verify by all users. A 50-50 split in user requirement was found for verification of more detailed parameters.
- Users considered verification of accompanying wind data as a high priority. Verification for high energy events and a separation of the verification according to wind-sea and swell dominated conditions were identified as important specific aspects of model performance to be tested.
- Quantitative measures of parameter errors were considered to be generally more important than measures of performance for predicting given events, with the exception of high energy storms.
- Where ensemble prediction system verification is conducted, users were keen to see performance cross-referenced against a deterministic forecast.
- Users expressed a preference to see verification statistics referenced against raw observations (i.e. without accounting for observation errors), a distinction made between in-situ and satellite data verification and an effort made to account for sampling and temporal variations within the verification's presentation.

[2] http://www.surveygizmo.com/s3/1306387/MyWave-Verification-Survey

- Metadata describing metrics, observed data used as a reference and quality control procedures should accompany the verification.

IV METRICS

IV.1 Identification of metrics by purpose

Verification is concerned with analysing the relationship between predictions of environmental conditions and their occurrence in reality, i.e. the joint probability distribution between prediction and reference

$$\Pr\{R, M\},$$

where R describes the sample of reference data and M the sample of predictions. Since the joint distribution is difficult to present and describe concisely, particularly for multi-dimensional data (Murphy, 1991), the majority of metrics are based on a simplification to a single scalar dimension. Different scalar measures assess a number of attributes of the prediction performance:

- Accuracy, i.e. a value representing overall quality of a set of predictions.
- Bias, a measure of any systematic difference between the sample of predictions and reference.
- Reliability, which describes accuracy and bias conditional on specific ranges of the predictions.
- Resolution/discrimination, which describe how well predictions in specific ranges are associated with similar sub ranges of the reference sample.
- Sharpness, which describes the prediction variability with respect to background climate, i.e. how much the prediction attempts to replicate the reference 'signal'.

In this document metrics are classified according to the purpose that each aims to fulfil. Clearly defining what each metric does is important to a MCS application since, in general, it is expected that users are unlikely to wish to review large numbers of metrics and will instead want to quickly discover those key pieces of verification data that meet a specific need. Five overarching purpose categories are identified according to the aspect of model performance being tested:

- Climatology tests (Annex A, Table C) determine the ability of the prediction system to replicate the reference climate, for example describing sharpness and bias of the predictions. These tests ignore any time-referencing in the sample pairs. The outcomes may be used to determine systematic errors and specific process representation issues.

- Measures of prediction uncertainty in parameter space (Annex A, Table M) estimate accuracy from the samples of prediction-reference errors. These metrics enable the errors to be viewed in context against background conditions or in prediction system intercomparison. The metrics may also be able to be used in estimating confidence limits for predictions or as information to underpin prediction correction strategies adopted by users.
- Measures of prediction accuracy and resolution in probability space (Annex A, Table P) describe the ability of the prediction system to successfully identify reference conditions, and are often used to determine the skill of a system against a baseline prediction. These data can also be used to evaluate the long term benefits of using the model predictions, i.e. whether more gains than losses will be made through basing decisions on prediction data.
- Measures of performance through the parameter range (Annex A, Table R) assess reliability and explore the impact of conditional corrections to the original predictions. These tests are described specifically in this document as a special case of data stratification.
- Extreme statistics (Annex A, Table X) analyse performance of the model specifically at the tail(s) of the distribution of conditions. The tests described are intended to be robust when working with limited data samples.

Metrics falling into each purpose category are identified within the following subsections. These describe (respectively) a set of core metrics that can be applied commonly to parameters observed by either in-situ or satellite remote sensed instruments (common metrics) and extensions to the metric set that can be achieved when specific properties of in-situ and satellite remote sensed reference data are taken into account. Tabular summaries of the full set of proposed metrics by purpose category are given in Annex A.

IV.2 Common metrics

IV.2.1 Role and limitations of common metrics

The envisaged role of metrics that can be applied regardless of the observation source is to provide a core set of common and widely available verification data that allow users to work with consistent uncertainty information for any region covered by MCS predictions. The aim is to make verification data available over the maximum geographic coverage using a mixed portfolio of observed references so that verification is available in both areas of the deep ocean where in-situ data are virtually non-existent or where satellite data quality is limited near to the coast. Where both data types are available the incentive of being able to compare verification and better estimate uncertainty from a combined data sample exists.

To achieve common ground between in-situ and satellite data (which have different sampling regimes, see Annex B, subsection VIII2.2) the sample of events verified must comprise instantaneous ‘snapshots’ of given parameters at specific locations/times (as opposed to, for example, a sample comprising event durations). The impact of viewing the match-up sample in this manner is that, strictly speaking, the metrics applied cannot assume or make use of any spatial or temporal linkage between events. In reality, if the sampling rate is high, such links will be present and it may then become important to ensure that the sample used is not aliased by particular sub-collections of data within the sample (e.g. Annex B, subsection VIII.4.1).

Within this subsection the description of metrics is focused on single parameters rather than application to multi-variate cases. The extension to multi-variate cases is discussed further in subsections IV.3, IV.4 and within Annex B, subsection VIII.2.3

IV.2.2 Common metrics for climatology tests (Table C)

One characteristic of the model’s representation of climate that cannot be reproduced using common metrics is how well the model reproduces parameter signal variations in the time dimension. The omission is necessitated by the assumption that no temporal links exist between sampled events.

If available, a long term climatology may provide a useful baseline predictor for these tests since, due to interannual and/or seasonal variability, a skilful prediction system might be expected to reproduce the short term climate better than a long term estimate.

Test C1: Reproduction of general features of the reference climate

Viewing a collection of parameter space metrics allows the user to quantitatively assess how well the parameter distribution has been reproduced and assess the sharpness of the predictor. The data also provide a useful background description of climate to accompany other tests and, because of this, it is proposed that the values provided are absolute quantities. The most concise metrics are based on comparing moments of the event sample distributions and should include higher moments of the distribution relating to skewness and kurtosis, since many parameters being tested (e.g. significant wave height, wind speed) cannot be assumed to be normally distributed.

Proposed metrics (in combination):

- Parameter mean, $\mathrm{E}[x] = \frac{\sum x}{n}$ (for variable x with sample size n); differentials in reference and predicted means measure bias
- Parameter root mean square (RMS) value, $\mathrm{RMS}[x] = \sqrt{\frac{\sum x^2}{n}}$
- Parameter standard deviation $\sigma = \sqrt{\mathrm{Var}[x]} = \sqrt{\frac{\sum (x - \mathrm{E}[x])^2}{n}}$
- Parameter skewness $\gamma = \mathrm{E}\left[\left(\frac{x - \mathrm{E}[x]}{\sigma}\right)^3\right]$
- Parameter kurtosis, $\beta = \mathrm{E}\left[\left(\frac{x - \mathrm{E}[x]}{\sigma}\right)^4\right]$, or kurtosis exceedence from the normal distribution value, i.e. β - 3

Test C2: Reproduction of details of the reference climate

Distribution comparisons can be used to provide more detail in representation of the reference climate and highlight sub-ranges of conditions which are particularly well or poorly replicated. For example, Quantile-quantile (Q-Q) plots view the samples by comparing values achieved at different percentiles within the cumulative probability distribution and provide a more useful visualization for the distribution tail than direct comparison of the probability distribution function against a set of parameter bins.

Proposed metrics:

- Q-Q plot, for parameters with long distribution tails (e.g. significant wave height) split over two levels to resolve body and tail of distribution
- Comparison/anomaly of occurrence probabilities for binned parameter ranges (e.g. 0.0<Hs<=1.0m, 1.0<Hs<2.0m etc.)

IV.2.3 Common metrics to measure prediction uncertainty in parameter space (Table M)

For common metrics persistence cannot be used as a baseline since consistent time indexing is not assumed in the event sample, but a baseline prediction can be used based on either a random sample from the observations or the mean value. Using the mean is recommended as this places error data in context against climatological variability of the wave field.

Test M1: Quantify the scale of errors

In parameter space, Root Mean Squared Error ($RMSE$) and Mean Absolute Error (MAE) are the most recognised metrics for overall error description. $RMSE$, which is a composition of bias and a measure of error scatter, is a particularly popular metric, but has been demonstrated to have drawbacks when comparing data with similar levels of performance (Mentaschi et al., 2013). As a result it is recommended that $RMSE$ is presented alongside a breakdown of contributions to the metric as described in Test M1a. Mentaschi et al. (2013) also discuss use of a corrected normalised indicator following Hanna and Heinold (1985), which mitigates issues with $RMSE$ by symmetrically normalising the squared error data using both prediction and reference values.

Proposed metrics:

- Mean Absolute Error, $MAE = \frac{\sum |EP|}{n}$, where EP denotes the sample of errors for prediction (M) and reference (R), ($EP_i = M_i - R_i$)
- Root Mean Squared Error (as for parameter RMS with EP as the input variable)

- Hanna and Heinold (1985) symmetrically normalised $RMSE$; $SNRMSE = \sqrt{\frac{\sum EP_i^2}{\sum M_i R_i}}$

Test M1a: Assess effects of prediction 'sharpness and reliability' on RMSE

Reviewing the contribution to $RMSE$ of prediction variability, correlation or bias is expected to be useful to model developers studying the overall effects of system changes. Mean Square Error (MSE) comprises bias and error variance contributions as

$$MSE = \mathrm{Var}[EP] + \mathrm{E}[EP]^2,$$

where error variance further breaks down as

$$\mathrm{Var}[EP] = \mathrm{Var}[R] + \mathrm{Var}[M] - 2\,\mathrm{Cov}[M,R].$$

MSE can be normalised by Var[R] (to give a skill score relative to a naïve predictor based on the reference mean). The normalised variance component is a form of (squared) Scatter Index (SI, which has also been defined in other forms by Bidlot et al., 1997; Ardhuin et al., 2007; Filipot and Ardhuin, 2012). Breaking down the $SI_{RVar}{}^2$ used here gives

$$SI_{RVar}{}^2 = 1.0 + \frac{\mathrm{Var}[M]}{\mathrm{Var}[R]} - 2\frac{\mathrm{Cov}[M,R]}{\mathrm{Var}[R]}$$

in which the third term can be re-written in terms of correlation and variance using

$$\frac{\mathrm{Cov}[M,R]}{\mathrm{Var}[R]} = \mathrm{Corr}[M,R]\sqrt{\frac{\mathrm{Var}[M]}{\mathrm{Var}[R]}}\,.$$

The normalised prediction variance and correlation can, respectively, be viewed as measures of the prediction systems' sharpness (i.e. how much the prediction attempts to replicate the reference 'signal') and reliability (i.e. whether the prediction is able to track the reference as it transitions through the range of conditions). In an ideal situation the normalised MSE will be reduced when both the normalised prediction variance and the correlation tend to 1.0 (so that $SI_{RVar}{}^2$ tends to 0.0), and when the bias part tends to 0.0. However the relationship for $SI_{RVar}{}^2$ is minimised when normalised prediction standard deviation is equal to the correlation value and therefore MSE will favour prediction systems with lower variance as correlation reduces. Mentaschi et al. (2013) also demonstrate dependence between SI and bias, such that SI is reduced in cases where the prediction has a negative bias. It can be argued that for wave prediction neither a reduction in forecast sharpness or a tendency to under-predict are desirable qualities, and so the MSE breakdown as described should help to indicate if reduced $RMSE$ scores have resulted from either of these effects. When many predictions are being compared the Taylor plot (Taylor, 2001) provides a useful visualization of the $SI_{RVar}{}^2$ breakdown.

Proposed metrics (in combination):

- MSE normalised by reference variance
- Bias normalised by reference variance
- (Squared) Scatter Index, $SI_{RVar}{}^2$
- Pearson Correlation
- Standard deviation of prediction normalised by reference standard deviation
- Taylor plot

Test M2: Quantify parameter uncertainty for the predictions

Moments of the error distribution provide a concise estimate of distribution characteristics and it is recommended that in addition to mean and standard deviation, skewness and kurtosis values are provided to indicate significant deviations from a normal form.

Proposed metrics (in combination):

- Error mean (bias)
- Error standard deviation
- Error skewness
- Error kurtosis exceedence

Test M3: Compare errors from two prediction systems

All the previously described tests can be intercompared. In addition, comparing errors from two prediction systems in Q-Q format may also be useful to model developers and forecasters, since the plot provides detail as to which aspects of the error sample are changed and can highlight if major differences are found in the error distribution tail(s). For operational verification of a single model it may be useful to provide a Q-Q plot which references against a standard pdf (e.g. a normal distribution) as the reference. Such a comparison would enable forecast users to assess value in applying error bars to a deterministic forecast by 'dressing' the forecast using a standard pdf.

Proposed metric:

- Q-Q plot comparing error samples

IV.2.4 Common metrics to measure of prediction uncertainty in probability space (Table P)

The basis for all probability space metrics is a test of whether predictions successfully identify the reference state within a given tolerance, i.e.

$$\Pr\{M_{correct}\} = \Pr\{R - tol < M < R + tol\},$$

$$\Pr\{M_{incorrect}\} = 1.0 - \Pr\{M_{correct}\}.$$

When the joint probability of correct prediction and the probability of predicting a given event are considered, results of the test can be broken down into four mutually exclusive states denoting correct and incorrect identification of the event and correct and incorrect rejection of the event. Common terms for correct identification, incorrect identification and incorrect rejection are *Hit*, *False Alarm* and *Miss* respectively. Often the data are presented in the form of a 'contingency table' as below:

	Event observed	Event not observed
Event predicted	*Hit*	*False Alarm*
Event not predicted	*Miss*	*Correct Rejection*

A wide array of options for these metrics are available, both in terms of the criteria applied to define a successful prediction and the credit given to successful, 'near miss' and unsuccessful predictions can be altered. In order not to overcomplicate the system, it is expected that the role of MCS verification should be to provide only a basic set of measures to indicate prediction uncertainty. It is proposed that scores directly indicating the probability of forecast success or failure provide the most understandable metrics. To enable users to build on these data specific to their own decision making processes it should be considered if the underlying match-up data can be published so that users can pick up and apply the data to their own specific scoring methods.

Metrics derived using a random chance selection from the reference data as the baseline prediction may provide useful context for the proposed scores.

Test P1: Quantify likelihood of predictions to fall outside prescribed tolerance

In probability space, simple accuracy/resolution metrics can be generated by expressing the probability of a correct or incorrect forecast based on errors falling within different threshold tolerances (e.g. +/- 0.1m, 0.25m, 0.5m, > 1m for significant wave height). It is proposed that data are presented in terms of 'the risk of seeing a parameter error larger than value x'.

Proposed metric:

- Percentage risk of error greater than predefined values

Test P2: Quantify ability to predict event x

From a basic contingency table for a predefined event and tolerance, accuracy metrics can be defined. If the contingency table itself is published numerous metrics can be generated by knowledgeable users, but it is proposed that the MCS verification scheme would also publish a small set of critical and accessible parameters. Initially these are identified as:

$FractionCorrect = \frac{Hits + CorrectRejections}{SampleSize}$, which quantifies the chance that predictions successfully identify both events and non-events.

$SuccessRatio = \frac{Hits}{Hits + FalseAlarms}$, which quantifies the chance that an event will occur if predicted.

$FalseAlarmRatio = 1 - SuccessRatio$, which quantifies the chance that an event will not occur if predicted.

$MissRatio = \frac{Misses}{Misses + CorrectRejections}$, which quantifies the chance of an event occurring if not predicted.

Proposed metrics:

- Contingency table for event
- Percentage scores for: *Fraction Correct*, *Success Ratio*, *False Alarm Ratio* and *Miss Ratio*

Test P3: Quantify long term benefit of decision making using predictions of event x

A simple cost-loss assessment of forecasts can be provided using the principal that a predicted event is associated with a cost (the same value is taken for a false alarm or a hit) and any miss is associated with a loss. This allows an Economic Value score to be generated against a varying cost-loss ratio (C/L in the range 0 to 1) since the cost of the prediction system will be

$$EV = C.(Hits + FalseAlarms) + L.Misses$$

Relative scores can be generated by referencing against a baseline prediction.

Carrasco et al. (2013) discuss application of a relative score, following Richardson (2000), that is constructed from costs associated with a situation in which no forecasts are available (in the case where action is never taken the cost will be $EV_c = L(Hits + Misses)$) and a perfect forecast (cost is $EV_{perfect} = C(Hits + Misses)$), so that Relative Economic Value:

$$REV = \frac{EV_c - EV_{EPS}}{EV_c - EV_{perfect}}.$$

Proposed metric:

- Relative Economic Value score

Test P4: Quantify effects of altering prediction threshold(s) for event x

An alternative use of categorical verification is to examine effects of corrections to prediction thresholds in order to optimise the performance of the prediction data. An example would be adjustment of the threshold criterion for prediction that significant wave height exceeds value x, as is made in the ‘alpha factor’ approach to de-risking forecasts taken by Det Norske Veritas (DNV) for marine operations assurance (DNV, 2011). Using the revised threshold allows new contingency tables to be constructed and statistics derived using Tests P2 and P3 to be generated and presented graphically. Effects on P2 metrics can be provided graphically to the user in the form of Relative Operating Characteristic curves (ROC, e.g. Mason, 1982; Buizza and Palmer, 1998) which compare *Probability of Detection* (chance of correctly forecasting an event) against *False Alarm Rate* (chance of forecasting an event that did not occur), or alternatively a Detection Error Trade-off curve (DET, *Miss Ratio* versus *False Alarm Rate*; Martin et al., 1997) for cases where users are more interested in ensuring that an event is not missed.

Proposed metric:

A required task will be to explore the usefulness of these statistics within the user consultation process.

Proposed metrics:

- Contingency table comparisons
- Percentage scores for: *Fraction Correct*, *Success Ratio*, *False Alarm Ratio* and *Miss Ratio*
- Relative Operating Characteristic plot
- Detection Error Trade-off plot
- *REV* comparison

IV.2.5 Common metrics to measure performance through event range (Table R)

Test R1: Quantify errors through predicted event sub-ranges

The simplest and most recognised way of visualising the full range of the error sample against conditions is to use a scatterplot. Scatterplot data are particularly useful in contextualising conditional error or fitted relationships and detecting outliers in the error sample. Extra value can be added if a plotting scheme is used that illustrates the density of data in sub-regions of the sample space (e.g. using contouring, a hexbin plot or overlaying binned probability values).

Parameter space measures of error, similar to those described for overall performance, can be applied to binned sub-samples of event conditions. It is expected that these data might be useful to modellers attempting to understand system errors and forecasters checking for deviations in the forecast errors from simple rules of thumb. In particular bias and standard deviation of errors enable simple visualization, although a box and whiskers approach might be perceived as clearer by some users.

Proposed metrics:

- Scatterplot (including density data)
- Mean and standard deviation of errors over sub-sample bins
- Box and whiskers plot for errors over sub-sample bins

Test R2: Test if a fitted relationship improves the predictions

Conditional relationships that aim to minimise errors can also be derived by generating a ‘best fit’ relationship between model and observations. Numerous methods to derive these relationships are available but, in lieu of user feedback, for simplicity it is suggested that MCS verification might only examine a linear fit between data. Adopting the simplest option raises a question as to whether an MCS should also make its underpinning match-up data available within the service, so that users seeking to perform other tests can do so.

Prediction system data should be used as the independent variable since the most useful information that can be derived for users is one that corrects the prediction toward the reference. For any fitted relationship (at least) revised overall RMSE and breakdown statistics (Tests M1 and M1a) should be provided for the fitted data in order to make a comparison with the original model sample values.

Proposed metrics:

- Linear fit relationship
- (At least) Tests M1 and M2 comparisons

IV.2.6 Common metrics to assess performance in extreme conditions (Table X)

Test X1: Test that reference extremes are reproduced by the prediction system

The simplest test of whether predictions have the necessary sharpness to reproduce extreme conditions is to use a Q-Q comparison based on the tail of the distribution (Test C1).

For very high percentiles it may be useful to annotate the data with numbers of events above the percentile in order to assess the extent to which sample size will affect the comparison.

Proposed metric:

- Q-Q plot for upper percentiles of distribution (95%ile and beyond)

Test X2: Test that events in the tails of the predicted and reference distributions are well correlated

To understand visually how data are matched in time and examine outliers, a scatterplot will be an effective visual metric. It is recommended to overlay event pairs identified by a predicted value exceeding the extreme event threshold (set in parameter or probability space) and pairs identified using reference value exceedence in order to understand the potential for missed events and false alarms (i.e. whether the resolution of the predictions is adequate).

Proposed metric:

- Scatterplot based on pairs identified by exceedence of 95%ile for both prediction and reference data

Test X3: Quantify prediction threshold effects on risk of a missed event, and impact on the number of false alarms

In this test, which is a special case of test P4, the risk of a missed event (*Miss Ratio*) is calculated against a moving parameter threshold. The trade-off in *False Alarm Rate* can be presented on the same DET plot for contrast. These data should be presented alongside occurrence statistics for the event.

Proposed metric:

- Miss Ratio and False Alarm Rate plotted against parameter for reference 95%ile threshold

- Detection Error Trade-off plot with threshold data as sample points

IV.3 Extension to metrics using in-situ reference data

Match-up data samples acquired from in-situ data sources enable common metrics to be extended in two regards. Large parts of the in-situ dataset should be continuous in time and, as described in Annex B Table B.1, a number of collocated parameters are available dependent on platform type.

Test C3: Reproduction of temporal variability in the reference climate

Use of autopectra or autocorrelation comparisons might provide useful information to model developers. However it is expected that processing requirements, necessary to ensure continuity and stationarity in the data and significance of the tests at key process timescales, would be difficult to implement within an operational system. More practical and accessible metrics may be generated based on evaluating the distribution of parameter 'windows' (i.e. periods of time when the parameter exceeds or falls below a critical threshold) over a time-series of data. When window length in time is considered as a parameter in its own right, metrics described in tests C1 and C2 can be applied.

Proposed metrics:

- Mean and standard deviation of window length for prescribed thresholds
- Q-Q comparison of window length for prescribed thresholds

Extension to M tests

Where in-situ data are available the use of persistence as a baseline prediction system can be applied.

Where directional data are available methods can be extended to an error in directional space as described in Annex B, subsection VIII.2.3.

Test M4: Describe parameter uncertainty for bi-variate predictions

Where collocated significant wave height, period and direction parameters are available the most accessible evaluation of bi-variate errors is via a scatterplot (with density information included). For data using a direction component approaching the parameter in vector form to provide a visual of zonal and meridional component errors is recommended.

Proposed metric:

- Scatterplot (including density data)

Extension to P tests

Where in-situ data are available the use of persistence as a baseline prediction system can be applied.

All tests listed in Table P can be applied to predictions of condition windows (e.g. if a 12 hour period with significant wave height less than 2m was predicted, did it occur?). Similarly a successful forecast can be determined based on conditions for several parameters being met. When producing multi-variate probability space metrics, presentation of accompanying data showing failure rates associated with individual parameters may also be useful.

Test X4: Test for prediction of an extreme event within a time window

Reviewing data in a time window around a reference extreme should enable the user to understand if the prediction system provides 'timely' (rather than exact) indications of extreme events. Output metrics are scatterplots of maximum parameter value within the given event window, and distribution of timing differentials. The metric relies on independence between events.

Proposed metrics:

- Scatterplot of maximum parameter values within event window
- Distribution of timing differentials between prediction and reference maxima

IV.4 Extension to metrics using satellite remote sensed reference data

Match-up data samples acquired against satellite observations enable common metrics to be extended since large parts of the satellite dataset should be continuous in space and, as described in Annex B Table B.1, a number of collocated parameters are available dependent on platform type. Spatial sampling also creates the possibility of providing mapped metric data, subject to sampling requirements being met (Annex B, subsection VIII.3.1).

Extension to M and P tests

Test M4 and tests in Table P can be applied for collocated satellite parameters as in Section IV.3.

Test X5: Test that predictions indicate the areal extent of an extreme event

Reviewing data in an along-track distance window around a reference extreme should enable the user to understand if the prediction system provides ‘area consistent’ indications (rather than exact prediction) of extreme events. Output metrics are scatterplots of maximum parameter values within the given event window (defined by up and down crossing of a parameter threshold) and comparison of event window along-track distances and maxima locations. The metric relies on independence between events, differentiating this test from Test M2.

Proposed metrics:

- Scatterplot of maximum parameter values within event window
- Scatterplot of along track event windows
- Distribution of along-track location differentials between prediction and reference maxima

V SUMMARY AND NEXT STEPS

It is believed that a successful MCS verification scheme will be based on provision of uncertainty and performance data that is relevant to a variety of marine users and which can be clearly portrayed and simply discovered. This document proposes a number of metrics that are believed to have utility in describing specific aspects of uncertainty and performance of operational wave models on a regularly updated basis, and associates a purpose with each metric. Establishing a defined purpose for the metrics is considered important for MCS portrayal of the verification since it will enable rapid discovery of relevant metrics by different user types.

The proposal of these metrics is influenced by a number of technical considerations (as discussed in Annex B). A major constraint on the metrics and sample period they represent is the availability of observed reference data, particularly for parameters other than significant wave height. In order to calculate statistics with a reasonable level of accuracy it is expected that at least 3 month samples of the most common observed parameters will be required for regional sea areas, and that this will need to be increased to 6 month or 12 month periods where multi-variate or extreme data are tested. In terms of portrayal of data, it is identified that methods may need to be introduced that help communicate variability in the statistics resulting from the conditions sampled, effects of sample size and observation errors.

Assumptions about both the user requirements for certain metrics and technical feasibility of implementation and portrayal will require testing. This will be carried out through the process of user consultation and evaluation of proposed metrics as discussed in Section III of this document.

Additions to this process will be made by integration of metrics for ensemble forecast data and assessing the steps that can be taken to evaluate or mitigate the effects of observation errors within the metrics. These subjects will be dealt with in accompanying reports from MyWave Work Package 4.

VI REFERENCES

Ardhuin, F., L. Bertotti, J.-R. Bidlot, L. Cavaleri, V. Filipetto, J.-M. Lefevre and P. Wittmann, 2007: Comparison of wind and wave measurements and models in the Western Mediterranean Sea, Ocean Eng. 34(3-4), 526-541

Bidlot J.-R., M.W. Holt, P.A. Wittmann, R. Lalbeharry and H.S. Chen, 1998: Towards a systematic verification of operational wave models. Proc Third Int. Symp. on ocean wave measurements and analysis, Waves97, Virginia Beach VA, American Society of Civil Engineers.

Buizza, R. and T.N. Palmer, 1998: Impact of ensemble size on ensemble prediction. Mon. Weather Rev., 126, 2503–2518

Carrasco A., Ø. Sætra and J.-R. Bidlot, 2013: Cost-loss analysis of calm weather windows. Journal of Operational Oceanography, Vol 6 No 1,17-22.

Det Norske Veritas (DNV), 2011: Offshore Standard, Marine Operations, General (DNV-OS-H101), DNV Oslo, 10.2011.

Filipot, J.F. and F. Ardhuin, 2012: A unified spectral parameterization for wave breaking: from the deep ocean to the surf zone. J. Geophys. Res., 117, C00J08, doi:10.1029/2011JC007784

Hanna, S. and D. Heinold, 1985: Development and application of a simple method for evaluating air quality. In: API Pub. No. 4409, Washington, DC, Washington, USA.

Lefèvre J-M and L. Aouf, 2012: Latest developments in wave data assimilation. Proceeding from the ECMWF Workshop on Ocean Waves, 25-27 June 2012. ECMWF, Reading, United Kingdom.

Martin, A. F., G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, 1997: The DET Curve in Assessment of Detection Task Performance. Proc. Eurospeech '97, Rhodes, Greece, September 1997, Vol. 4, pp. 1899–1903.

Mason, I., 1982: A model for assessment of weather forecasts. Aust. Meteorol. Mag., 30, 291–303.

L. Mentaschi, 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

Murphy, A.H., 1991: Forecast verification: its complexity and dimensionality. Mon. Weather Rev., 119, 1590-1601.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. Quart. J. Royal Met. Soc., 126, 649-667.

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.

VII ANNEX A: TABLES OF PROPOSED METRICS

Tables in this section summarise metrics that are proposed to be tested within the MyWave user consultation on verification as part of WP4. The heading 'Target parameter(s) for user review' denotes data that will be used in testing of the metrics. The heading 'Extension' denotes parameters that could also be considered for verification within an operational scheme subject to observed data availability.

Table C: Climatology tests

Purpose	Primary Reference	Proposed Metric(s)	Target parameter(s) for user review	Extension
C1: Reproduction of general features of the reference climate	Common	Combined comparison of parameter: Mean; RMS; Standard Deviation; Skewness; Kurtosis Exceedence	Hs Tp/Tz for in-situ reference	Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data
C2: Reproduction of details of the reference climate	Common	Q-Q plot Binned probability plot and anomaly	Hs Tp/Tz for in-situ reference Binned probability anomaly for Hs-T	Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data Hs-Dirn for spectral in-situ and (sub-range of spectrum) SAR
C3: Reproduction of temporal variability in the reference climate	In-situ	Mean; Standard Deviation Binned probability plot and anomaly	Windows defined by Hs threshold	

Table M: Measures of prediction uncertainty in parameter space

Purpose	Primary Reference	Proposed Metric(s)	Target parameter(s) for user review	Extension
M1: Quantify the scale of errors	Common	MAE RMSE SNRMSE Baseline prediction is mean reference value	Hs Error Tp/Tz Error for in-situ reference	Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data Baseline reference can be extended to persistence for in-situ data
M1a: Assess effects of prediction 'sharpness and reliability' on RMSE	Common	Combined: Normalised MSE; Normalised Bias; (Squared) SI; Pearson Correlation; Normalised Standard Deviation Taylor Plot	Hs Error Tp/Tz Error for in-situ reference	Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data
M2: Quantify parameter uncertainty from for the predictions	Common	Combined: Mean; Standard Deviation; Skewness; Kurtosis Exceedence	Hs Error	Tp/Tz for in-situ reference; Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data
M3: Compare errors from two prediction systems	Common	Q-Q plot	Hs Error	Tp/Tz errors for in-situ reference; Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data
M4: Describe parameter uncertainty for bi-variate predictions	In-situ; Satellite	Scatterplot (including density information)	Hs-Tp/Tz Errors for in-situ Hs-MDir Errors for spectral in-situ	Hs-MDir for spectral in-situ and (sub-range of spectrum) SAR

Table P: Measures of prediction uncertainty in probability space

Purpose	Primary Reference	Proposed Metric(s)	Target parameter(s) for user review	Extension
P1: Quantify likelihood of predictions to fall outside prescribed tolerance	Common	% risk of error greater than predefined value Baseline prediction is mean reference value	Hs	Tp/Tz for in-situ reference; Spectral Hs and Mdir for spectral in-situ buoys and (sub-range of spectrum) SAR data Assess for single parameters and multi-variate Baseline reference can be extended to persistence for in-situ data
P2: Quantify ability to predict event x	Common	Contingency Table Combined % scores: Fraction Correct; Success Ratio; False Alarm Ratio; Miss Ratio Baseline prediction is mean reference value	Hs Windows for in-situ	Spectral sub-range Hs for SAR data Multi-variate for in-situ Baseline reference can be extended to persistence for in-situ data
P3: Quantify long term benefit of decision making using predictions of event x	Common	REV Score Baseline prediction is mean reference value	Hs Windows for in-situ	Spectral sub-range Hs for SAR data Multi-variate for in-situ Baseline reference can be extended to persistence for in-situ data
P4: Quantify effects of altering prediction threshold(s) for event x	Common	Tests P2, P3 ROC, DET curve	Hs	Extension to windows for in-situ

Table R: Measures of performance through event range

Purpose	Primary Reference	Proposed Metric(s)	Target parameter(s) for user review	Extension
R1: Quantify errors through predicted event sub-ranges	Common	Scatterplot (including density data) Sub-range bin error Mean and Standard Deviation Sub-range bin error box and whiskers	Hs	
R2: Test if fitted relationship improves the predictions	Common	Combined: Linear fit relationship; Tests M1, M2	Hs	

Table X: Assessment of performance in extreme conditions

Purpose	Primary Reference	Proposed Metric(s)	Target parameter(s) for user review	Extension
X1: Test that reference extremes are reproduced by the prediction system	Common	Q-Q plot above 95%ile	Hs above 95%ile	
X2: Test that events in the tails of model and observed distributions are well correlated	Common	Scatterplot	Hs above 95%ile (identified in both prediction and reference data)	
X3: Quantify prediction threshold effects on risk of a missed event and impact on the number of false alarms	Common	Miss Ratio and False Alarm Rate plotted vs parameter range Detection Error Tradeoff plot	Hs	
X4: Test for prediction of an extreme event within a time window	In-situ	Scatterplot Probability distribution	Hs maximum in window Hs maximum timing differential	
X5: Test that predictions indicate the areal extent of an extreme event	Satellite	Scatterplot Scatterplot Probability distribution	Hs maximum in window Event window lengthscale Hs maximum position differential	

VIII ANNEX B: TECHNICAL CONSIDERATIONS FOR WAVE VERIFICATION

The purpose of this annex is to highlight and discuss technical considerations relating to both the metrics proposed in Section IV of this document and requirements for underpinning reference data.

VIII.1 Background to verification metrics

Verification is concerned with analysing the relationship between predictions of environmental conditions and their occurrence in reality, i.e. the joint probability distribution between the prediction and reference

$$\Pr\{R, M\},$$

where R describes the sample of reference data and M the sample of predictions. Since the joint distribution is difficult to present and describe concisely, particularly for multi-dimensional data (Murphy, 1991), the majority of metrics are based on a simplification to a single scalar dimension. Different scalar measures assess a number of attributes of the prediction performance:

- Accuracy, i.e. a value representing overall quality of a set of predictions.
- Bias, a measure of any systematic difference between the sample of predictions and reference.
- Reliability, which describes accuracy and bias conditional on specific ranges of the predictions.
- Resolution/discrimination, which describe how well predictions in specific ranges are associated with similar sub ranges of the reference sample.
- Sharpness, which describes the prediction variability with respect to background climate, i.e. how much the prediction attempts to replicate the reference ‘signal’.

The usual approach is to assess these attributes by analysing the sample distribution of prediction–reference errors EP defined by the difference between matched pairs of prediction and reference data

$$EP_i = M_i - R_i .$$

A generalised form for an error pair, where the reference is an observation, is illustrated in Figure B.1. This form is applicable to verification of both probabilistic and deterministic predictions, and cases which aim to account for observation errors. In the figure both prediction and reference values for a given parameter space (which for simplicity has been shown in a single dimension, but could be multi-dimensional or even circular) have uncertainties associated with them, which are shown in the form of pdfs. For the predicted data this form would be adopted if some measure of uncertainty were being applied to the predicted data (e.g. by using an ensemble prediction system). In the deterministic case the pdf is simplified to having a value of 1 at the predicted parameter value and zero elsewhere. The reference data sample will be drawn from the joint probability distribution of the true condition plus an observation error.

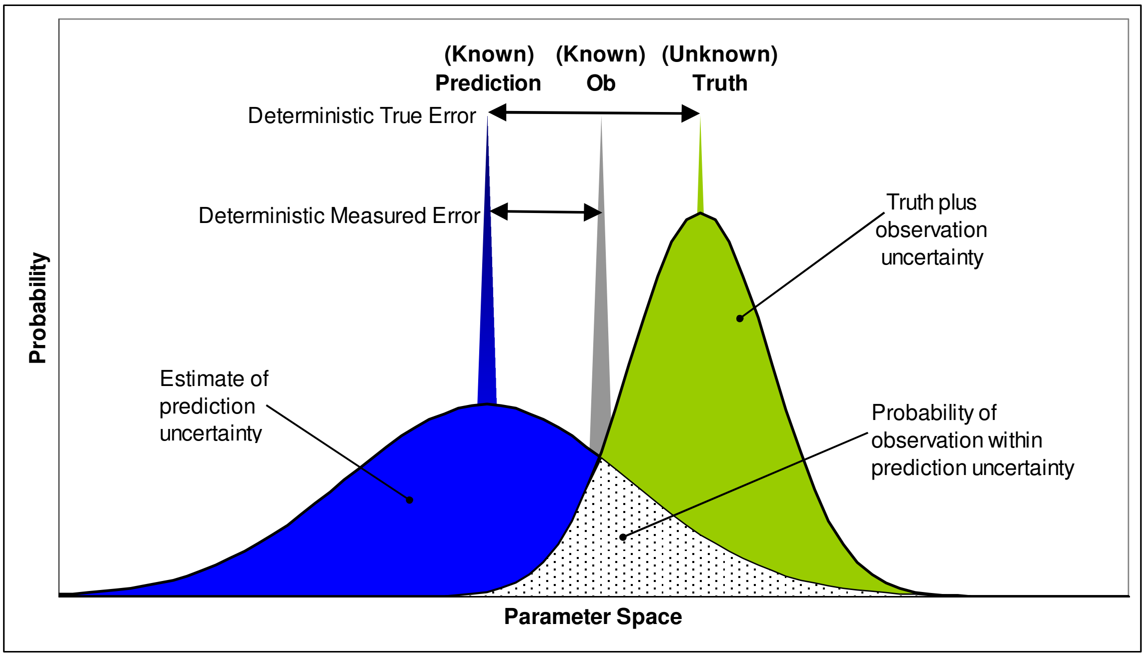
(Known)
Prediction
(Known)
Ob
(Unknown)
Truth
Deterministic True Error
Deterministic Measured Error
Truth plus observation uncertainty
Probability
Estimate of prediction uncertainty
Probability of observation within prediction uncertainty
Parameter Space

Figure B.1. Schematic for parameter/probability space definition of prediction-observation error.

In this form two approaches to deterministic prediction verification can be taken. The first is to examine the sample estimate of the pdf of parameter space distances (errors) between the prediction and reference. In theory this estimate can be used to quantitatively inform uncertainty estimates in working predictions (e.g. operational forecasts) or as a measure of the impact of changes to the prediction system. Metrics associated with this approach define the accuracy, bias, reliability and sharpness of the predictions.

The second approach is to quantify probability of the prediction correctly or incorrectly identifying the reference state, based on prediction and/or reference pdfs and some predefined parameter space tolerance. In the case of deterministic verification the assumed pdfs are functions that have probability of 1.0 at the predicted and reference values and 0.0 elsewhere. The probabilities generated define the uncertainty associated with using the prediction system in its original form for decision making. Often the metrics focus on a specific event and correct/incorrect data are further subdivided into identification/rejection categories. The metrics tell the user about resolution and discrimination properties of the predictions in addition to accuracy.

For the purpose of this and other WP4 documents the two approaches will be referred to as ‘parameter space verification’ and ‘probability space verification’ respectively. It is expected that the different outcomes from parameter space and probability space metrics will resonate differently with different user types, and one task of WP4 will be to identify whether this is the case.

VIII.2 Definition and availability of wave parameters

VIII.2.1 Definition of wave data and consequences for verification

In the context of wave prediction and observing, the term ‘wave’ is a catchall for statistics describing the population of individual waves propagating over the ocean surface. Predictions and observations will vary in the detail at which these statistics are estimated and include:

- Significant wave height, period, direction and spreading parameters estimated from a weighted summation of energy in the full two-dimensional wave spectrum
- Similar parameters defined using a predefined sub-range of the spectrum
- Details of energy and direction information over the spectra
- Details of energy distribution over the (2D) frequency-direction spectrum

This leads to a situation where metrics can be applied to number of wave statistics independently. However, in order to review the true overall performance of a wave model, understanding how particular parameter combinations are predicted is also important. Common to weather model verification, for any increase in the detail at which predictions are reviewed there will be a trade-off in terms of a) the required data sample size that will enable statistics to be robust; b) in being able to draw simple and useable conclusions from the data. Model developers assessing detailed performance changes over long time periods should seriously consider the data that can be obtained from metrics derived using wave spectra (Bidlot et al., 2005). However, for an operational scheme which is particularly targeted at downstream forecast agencies and lay-users, concise and conclusive metrics are more likely to be generated based on summary parameters derived from either the full spectrum or some well defined and practically useful sub-ranges.

VIII.2.2 Available observations of wave parameters

Although availability of data has significantly improved in the last 20 years, wave observations are still sufficiently sparse to be a limiting factor in the verification that can be practically generated. This is particularly the case for operational verification that generally uses data sampled over periods of a few months. Here two observed sources of reference data are focused on. 'In-situ data' describes any form of observation (e.g. using a heave sensor, laser altimeter) made from platforms that are fixed in space and sample at regular short intervals in time. 'Satellite data' describes remote sensed observations made by instruments mounted on low orbit space vehicles. These platforms are not geostationary and so the observations are made along tracks following the satellite's (polar) orbit of the earth. This leads to a data sample that is spatially dense along-track but temporally sparse at fixed points.

Wave parameters commonly observed by various instruments are listed in Table B.1. What becomes immediately apparent is that significant wave height and wind speed are observed in significantly higher volumes than other wave data. This may lead to a requirement for different sample periods to be used for verification of different parameters.

VIII.2.3 Treatment of circular and vector parameters

As defined in subsection VIII.2.1 the full wave field or its components are ideally represented as (at least) a 3-dimensional entity comprising the energy associated with the waves (usually expressed as significant wave height), their period (which also relates to speed of propagation) and the direction in which the energy is transmitted. These characteristics are not independent and a verification scheme should consider what degree of multi-variate testing should be carried out. In addition, direction is a 'circular' variable and this needs to be dealt with in data processing.

Table B.1. Availability of observed parameters for wave verification

Wave Parameter	Available Platforms	Notes
Significant wave height	In-situ (approx. 400 instruments globally) Satellite Altimeter (generally 2 missions available)	Mix of instrument types
Peak wave period	In-situ (approx. 270 instruments globally)	Mix of instrument types
Mean zero-upcrossing wave period	In-situ (approx. 150 instruments globally)	Mix of instrument types
Mean/peak wave direction	In-situ (approx. 150 instruments globally)	Data from spectral sensors
Mean/peak wave directional spread	In-situ (approx. 150 instruments globally)	Data from spectral sensors
Frequency range wave energy	In-situ (approx. 150 instruments globally) SAR (generally 1 mission available)	Data from spectral sensors Spectral sub-range available
Frequency range wave direction	In-situ (approx. 150 instruments globally) SAR (generally 1 mission available)	Data from spectral sensors Spectral sub-range available
Sub-range wave height	In-situ (approx. 150 instruments globally) SAR (generally 1 mission available)	Data from spectral sensors Spectral sub-range available
Sub-range wave direction	In-situ (approx. 150 instruments globally) SAR (generally 1 mission available)	Data from spectral sensors Spectral sub-range available
Maximum wave height	In-situ (approx. 20 instruments globally)	

Regarding the direction parameter, the majority of the metrics described in this document can be applied if the error between directions is considered simply as a clockwise or anti-clockwise shift. The shift chosen is taken as the minimum rotation required, e.g. for the case of direction expressed in degrees

$$D_{shift} = \min\left(\left|D_M - D_R\right|, 360 - \left|D_M - D_R\right|\right).\mathrm{sgn}\left(D_M - D_R\right).$$

Scatterplots of direction data (and hence fitted relationships) can also be generated by extending axes above and below 0 and 360 degrees respectively and applying the shift value to the independent (x-axis) variable.

Achieving clarity in the multi-variate case is simple for probability space metrics, since these test events rather than measuring differentials, but is more complex for parameter space metrics. Based on this requirement and the types of data that would be regularly available for operational verification, bi-variate analyses based on the following parameter pairs are recommended:

- Wave height error and direction shift
- Meridional and zonal wave energy error (i.e. how much wave energy is transported north and east)
- Wave height error and period error
- Wave group speed and direction shift.

At the point of consultation with users, it is expected that assessing the utility of graphically expressing the distribution of bi-variate errors (e.g. via a scatterplot with axes in parameter error space) is required before approaching any statistics where a distribution might be fitted to the data.

VIII.3 Sampling requirements

VIII.3.1 Estimate of sample size requirements

In order to place the impact on metrics of sample size in context, two simple cases can be looked at. In parameter space the mean of variable x will, assuming the central limit theorem, follow a normal distribution with variance σ and for sample size n can therefore be determined to within a given sample space 95% confidence interval using

$$\bar{x} - \frac{2\sigma}{\sqrt{n}} < CI < \bar{x} + \frac{2\sigma}{\sqrt{n}}.$$

For an example of the northern North Sea, where winter error standard deviation for significant wave height can be in the order of 0.6m, a confidence interval of +/-2cm for the mean error would be obtained from a sample of 3600, and an interval of +/-5cm would be obtained from a sample of 576 data points.

In probability space a proportion of successful or unsuccessful forecasts will be estimated. In the case of independent data the estimate should follow a (scaled) Binomial distribution and has a maximum variance of 0.25n when the proportion value is 0.5. Following the Wald (1943) method for the binomial distribution and using the maximum variance value the sample size for a 95% confidence interval of given width (W) can than be obtained using:

$$n = \frac{4}{W^2}.$$

Thus for a confidence interval for normalised probability of +/-0.05 (10%) a sample of size 400 is required and for an interval of +/-0.01 (2%) a sample size of 10000 is needed.

From these rough estimates it is clear that minimum independent samples of the order of between 400-600 points are desirable for wave verification.

VIII.3.2 Use of independent data

The metrics in this document assume use of a sample of independent events. This assumption becomes a necessity in order to understand the confidence that can be placed in these statistics (for example to assess the effects of sample size using re-sampling methods as recommended in a recent update to World Meteorological Organisation from the Coordination Group for Forecast Verification, 2012). Work by Greenslade and Young (2005), Janssen et al (2007), and Palmer and Saulter (2013) suggests that wave data are well correlated over time and space scales in the range 12-18 hours and 50-300km, although these quantities are regionally and parameter dependent (e.g. longer scales exist for significant wave height in swell dominated tropical areas than in higher latitude storm tracks).

Figures B2 and B3 assess the effects of applying independence criteria to samples of significant wave height obtained from in-situ and remote sensed observations in the North Sea. This was achieved by applying temporal and spatial restrictions of 12 hours and 220km following Palmer and Saulter (2013). The sub-sampling scheme used to generate the examples applies both criteria using a ‘first come first served’ approach, i.e. once a reference data value has been read which occupies a given space-time coordinate any subsequent values within a set distance-time range are rejected. To check how fairly the scheme works the order of the data was randomised before sub-sampling and the process was repeated 50

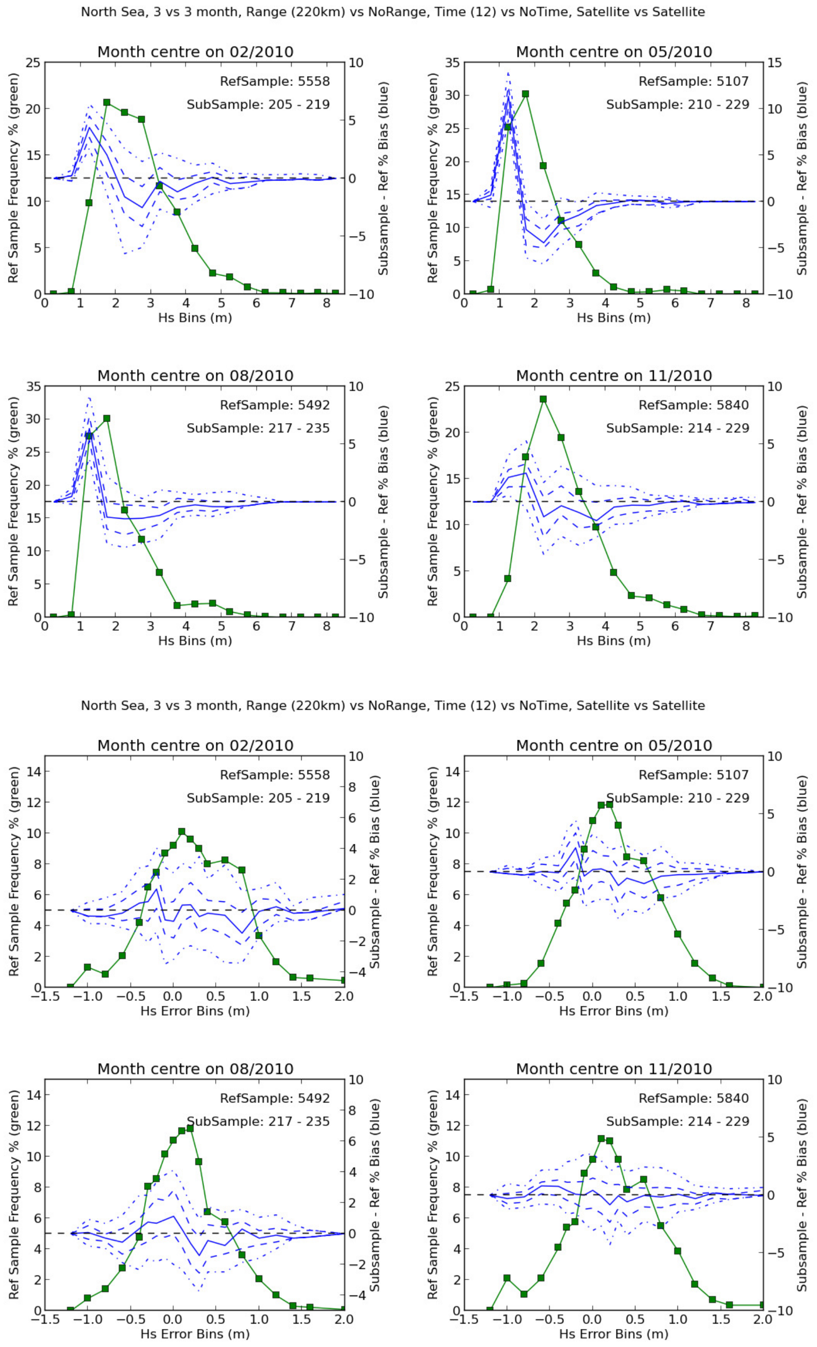

Figure B.2. Effects of using a sub-sampling technique to achieve an independent sample of satellite (Envisat) data in the North Sea. The top four panels show 3 monthly samples of observed significant wave height, and the lower four panels show samples of model-observation error for the same periods. The green line indicates the original observed distribution, whilst the blue lines show the median (solid), 25th and 75th percentile (dashed), and 5th and 95th percentile differential in sample distribution for 50 sub-samples.

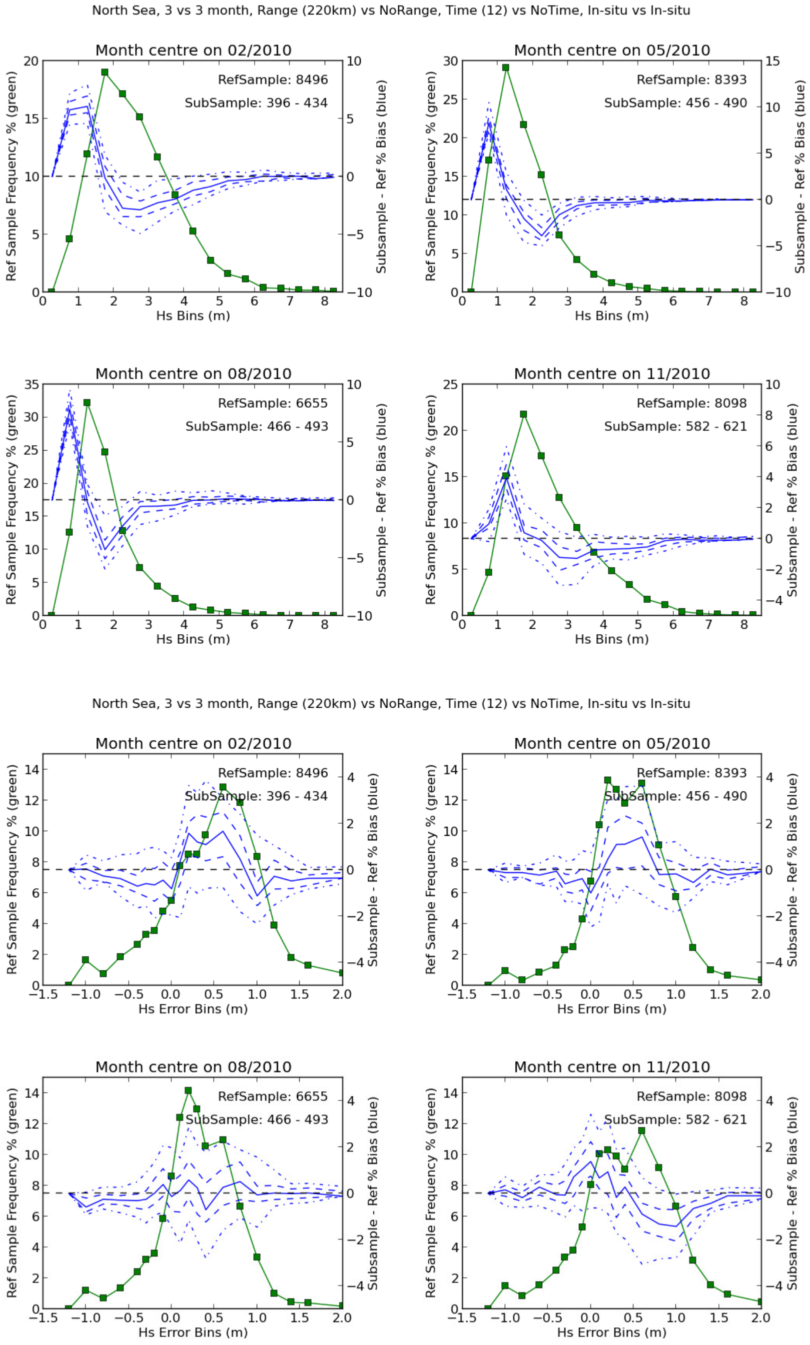

Figure B.3. Effects of using a sub-sampling technique to achieve an independent sample of in-situ data in the North Sea. The top four panels show 3 monthly samples of observed significant wave height, and the lower four panels show samples of model-observation error for the same periods. The green line indicates the original observed distribution, whilst the blue lines show the median (solid), 25th and 75th percentile (dashed), and 5th and 95th percentile differential in sample distribution for 50 sub-samples.

times – hence the changes to the sub-sampled data are shown as a plume in the figures rather than a single line. The principle result of this analysis is the reduction in data sample size resulting from adopting these strict sampling criteria. In both figures data volumes are reduced by close to a factor of 20.

The North Sea is expected to be representative of other regional seas in terms of observed data coverage, if not better populated. So the analyses presented in Figures B2 and B3 suggest that for verification focused on a limited spatial area a 3 month sample period is likely to be the minimum required to achieve a robust independent sample of data. The sampling period should be correspondingly increased if focusing on specific parts of the data sample or when evaluating combined parameters. For example, a contingency table defined for event prediction verification stratifies the verification sample according to prediction or non-prediction of the event of interest. In this case the confidence levels associated with the results need to be judged based on the sample size in either predicted (for hits and false alarms) and non-predicted categories. In the example of a rare event the hit and false alarm proportions may have much larger confidence intervals placed upon them than the miss and correct rejection proportions due to the relative difference in the sample sizes.

VIII.4 Additional considerations for metric generation

VIII.4.1 Conditional influences on the metrics

A consideration for operational metrics that are updated regularly and use short sample periods is how to communicate conditional influences on the statistics associated with the samples used. The requirement to do this is driven by potentially different applications of verification data by different users. Traditionally the model developer view of verification is either as a method of reviewing recent performance of a model in order to identify system issues, or as data with which to make intercomparisons between systems using a consistent baseline. A forecaster or decision making user may wish to use the data differently however, for example by applying measures of uncertainty to future predictions in order to estimate decision risk. This second example, where the verification data are applied downstream, might be particularly sensitive to conditional effects.

For example, Figure B4 shows the relationship between root mean squared error (RMSE) and both mean significant wave height and standard deviation of wave height (as proxies for the distribution of conditions) in monthly reference samples for 1 year of data from North Sea

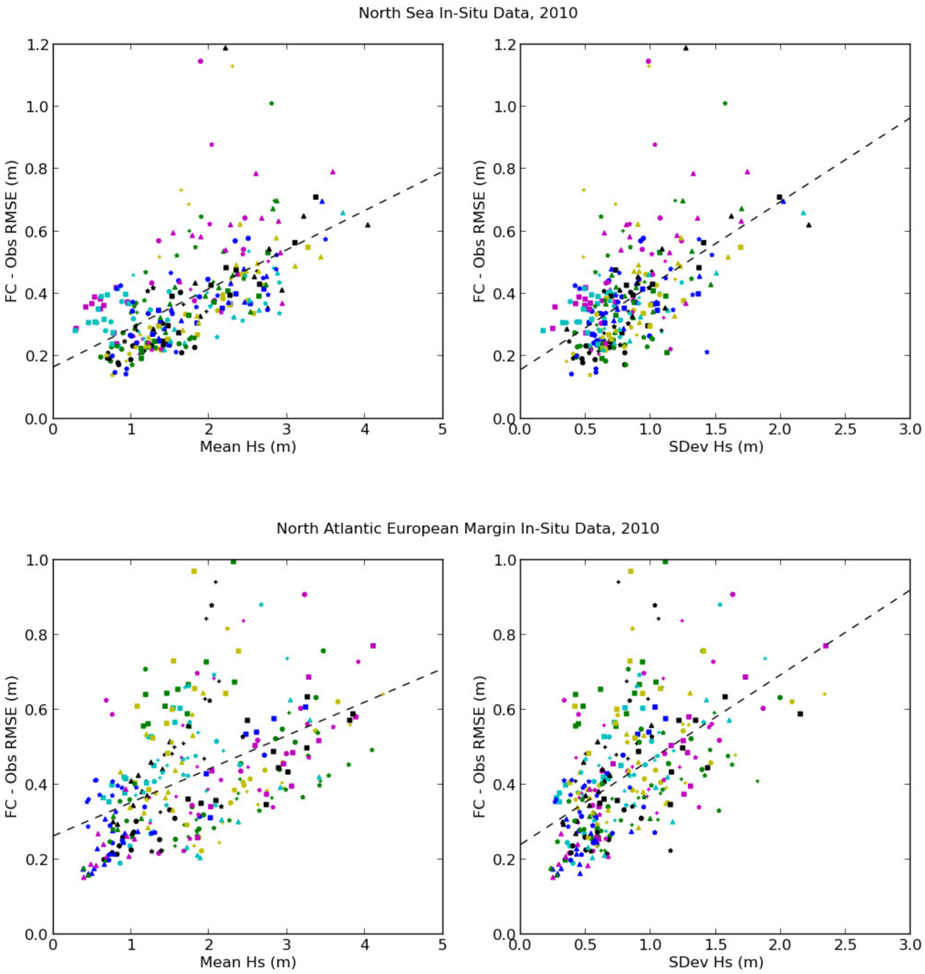

Figure B4. Monthly values during 2010 of mean and standard deviation of significant wave height versus RMSE for in-situ platforms in a) North Sea, b) North Atlantic European Margin regions. Colours and shapes denote different platforms.

and North Atlantic in-situ platforms. The data are well scattered from site to site and month to month, and a trend for increasing RMSE with both parameters is present. Other metrics may be more sensitive, for example threshold based contingency statistics for events with low sample size. In Figures B2 and B3 the distribution of observed parameter values were changed significantly when independence criteria were applied, suggesting some aliasing of data in the original sample toward particular conditions. Aliasing is likely to have occurred in the in-situ sample due to clustering of observations from oil platforms in particular areas around the North Sea, and can also be sensibly explained for the satellite by considering the sample of fetch lengths that can be achieved for various locations in the North Sea. The satellite will sample regularly from central locations (with similar fetches onto site of the order 200km or more) and less regularly from coastal locations (for which some sectors will see extremely short fetches). The impact on the error sample in the figures appears much more limited.

The simplest approach to mitigating this issue is to present the verification as sampled, and with appropriate caveats. This is the approach that has been adopted up to now within MyOcean (Alastair Sellar, *pers. comm.*). In this instance it is then the user's decision as to whether and how strictly to apply these data as part of a downstream process. Variability in these metrics might be reinforced to users by presenting longer term records of the metrics.

However, it may be possible to explore other methods. For example drawing data from the original verification sample in order to more consistently represent the event population associated with a long-term regional climatology of conditions. The verification described in this instance should be representative of a sample from the population of errors that a user might experience if operating throughout the region and over the long term. An alternative is to generate a revised sample or metrics in which reliability associated with events of different types is equally represented, i.e. the final statistics should be (reasonably) independent of the underlying climatology. Data conforming to this ideal can be established based on a stratification of the sample, for example using Neyman allocation in which the revised (optimal) sample is drawn from strata according to the local sample estimate of standard deviation of errors. For long tailed distributions typical of wave parameters, using preset condition ranges (e.g. 0-1m, 1-2m etc. for significant wave height) is not practical, but forms of sampling where either the original sample is broken into a small number of equal width percentile sub-ranges, or following a Cumulative Frequency of the Square Root method (Dalenius and Hodges, 1959) should be acceptable. The trade-off with both re-sampling techniques is that the level of complexity for the metric is increased whilst still not entirely meeting the user need. Identifying if any of these methods are applicable to user focused verification could be included within the project consultation if time is available.

VIII.4.2 Application of resampling techniques

Consistent with recommendations from the WMO Coordination Group for Forecast Verification (2012), application of resampling techniques, e.g. using the Bootstrap (Efron and Gong, 1983), in order to assess sensitivity of the metrics should be considered by MCS verification. The approach is not without compromise however. Analyses of resampled metrics introduce an extra level of complexity to presentation and interpretation of verification data. In addition processing requirements for the data may be significant. A suggested method to reduce the processing overhead is to employ a Block Bootstrap (Carlstein, 1986; Kunsch, 1989). The method maximises use of an original (dependent) data sample by dividing the sample into blocks of sufficient size that can be considered as an independent subset of the overall data sample. In this instance the identification of independent data blocks needs only to be made once and the need for multiple simulations of observation errors is also reduced as the sample size for each bootstrap member is maximised. The main issue is in adopting a best method for establishing the data blocks and ensuring that each supplies an equal number of observations to the bootstrap member samples. This is likely to be more of a concern for in-situ data than for satellite data due to the clustering of in-situ platform locations.

Figure B4 shows results from application of a scheme that used sampling in predefined areas and time periods as the basis for block selection for both in-situ and satellite data. In these tests a limited area of the northern part of the North Sea was divided into non-overlapping 2 degree latitude by 4 degree longitude boxes in order to define the area blocks and non-overlapping 24 hour periods to define the time blocks. The area blocks were set in order to ensure that an even number sample of observations would be captured in each area, however the block resampling scheme also identified a minimum sample size for each block and rejected or sub-sampled from the blocks based on this value. Data volumes used in the statistics were reduced compared to the original sample by a factor of 2-4, rather than the factor of 20 described for the independent data test in subsection VIII.3.2. In Figure B5 it is demonstrated that whilst the sample mean of predicted conditions used in the verification did not vary significantly between instruments (with the exception of the final quarter of 2010), the MAE values generated are distinct and suggest a possible discrepancy between the sampled instrument measurements of conditions.

Whilst the example in Figure B5 allows us to infer the trust we can place in the statistics for particular examples, questions need to be asked as to whether general application of this type of procedure can be made to work in a wide ranging operational verification scheme, and whether the resulting presentation of data will be suitably accessible for users.

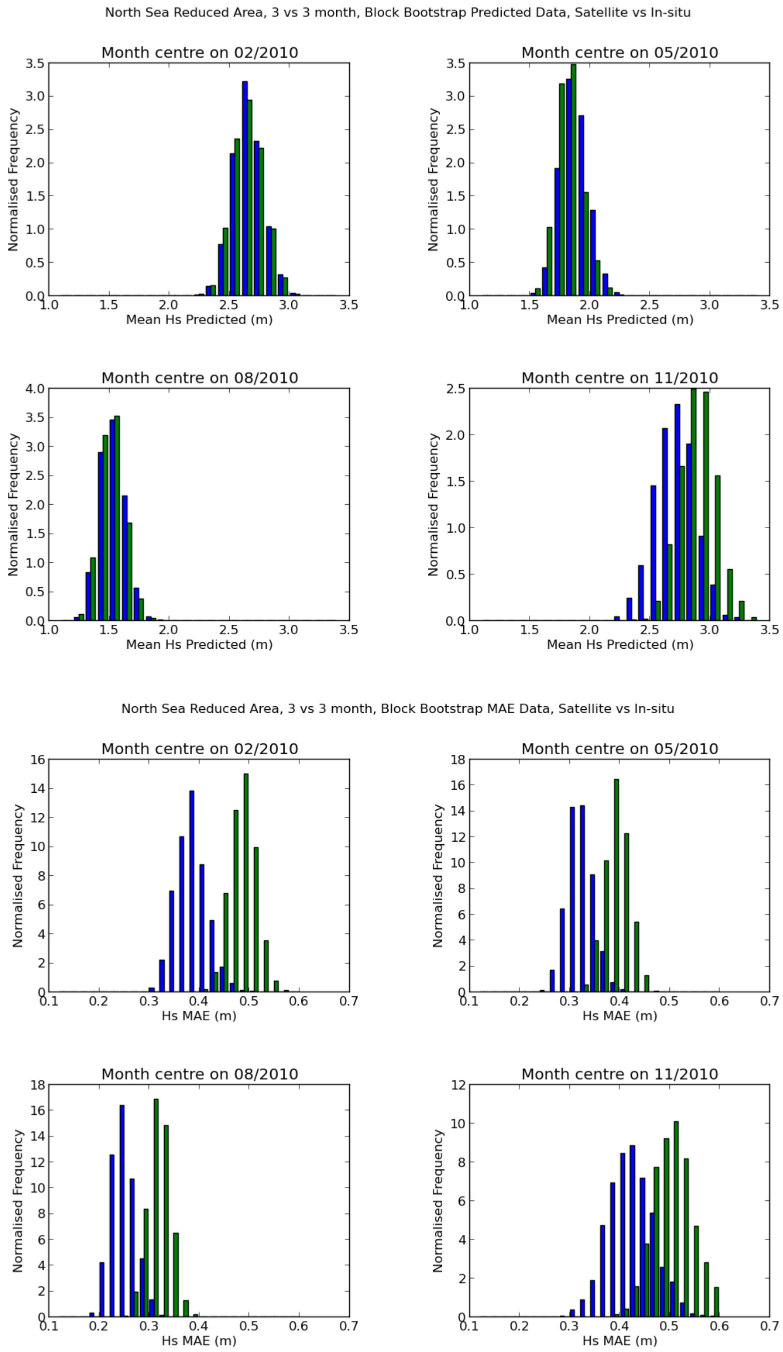

Figure B5. Comparisons of model sample mean and model-observation MAE values for 3 month block-bootstrap samples of in-situ (blue) and Envisat altimeter (green) significant wave height match-up data.

VIII.4.3 Accounting for observation errors

A final consideration for a verification scheme that may reference against different observed sources is that each observation type will have different error characteristics. Two types of observed errors can be considered, namely 'representation errors' and 'instrument errors'. Within a verification scheme 'representation errors', which relate to the scales over which waves are sampled by the observing system or represented in a model, can be mitigated by standardising the scales over which observations are aggregated as much as possible. Achieving consistency in this respect is desirable as it means that the role of representation errors in the verification will only vary with the model used. Previous work has used an estimated model scale to determine aggregation schemes for observations (e.g. Bidlot and Holt, 2006; Bidlot et al., 2007). However, for a MCS verification system which potentially compares different models, and bearing in mind the common perception that in-situ observations are the de-facto standard for true sea-state, it can be argued that referencing against a 15-30 minute in-situ sample enables more consistent and user focused intercomparison. This length of sample at a fixed point is equivalent to an approximate 8-25 km area sample for deep water wave energy at period 5-16 seconds and, for example, would be represented by 2-4 data points from altimeter soundings at 1Hz frequency.

Instrument errors comprise systematic and random components. Evaluating (relative) systematic and random errors within the observations requires assessment of the observations over a long time period (e.g. through triple collocation studies; Janssen et al., 2007), particularly if errors through the parameter sub-range are to be assessed (e.g. Abdalla et al., 2010). Incorporation of observed error estimates within ensemble prediction verification scheme has been demonstrated by Saetra et al. (2004). An approach, based on Saetra et al. (2004)'s method, and applicable to deterministic model verification will be discussed further in an accompanying MyWave report as part of deliverable D.4.3. However it is important at this stage in the consultation to pose the question as to whether verification against an estimated truth is more relevant to users than direct comparisons of the consistency between predictions and observations, since the latter will more likely reflect experience of direct use of observations and predictions in the operating environment.

VIII.5 Summary

The aim of this annex has been to identify and discuss technical issues that influence the proposed verification metrics discussed in Section IV and their method of portrayal. A major constraint on the metrics and sample period they represent is the availability of observed reference data, particularly for parameters beyond significant wave height. In order to calculate statistics with a reasonable level of accuracy it is expected that at least 3 month samples of the most common observed parameters will be required for regional sea areas, and that this will need to be increased to 6 month or 12 month periods where multi-variate or extreme data are tested.

In terms of portrayal of data, it is identified that methods may need to be introduced in order to help communicate variability in the statistics that may be introduced by the conditions sampled, effects of sample size and observation errors. In reviewing whether and how to implement these, the project will need to consult with users and be mindful that accessibility and simplicity in the metrics will be of paramount importance to a successful MCS verification scheme.

VIII.6 References

Abdalla, S., P.A.E.M. Janssen and J.-R. Bidlot, 2010: Jason-2 OGDR Wind and Wave Products: Random Error Estimation. ECMWF Technical Memorandum No. 639.

Bidlot, J.-R., P.A.E.M. Janssen and S. Abdalla, 2005: On the importance of spectral wave observations in the continued development of global wave models. Proc. 5th International Symposium WAVES 2005.

Bidlot, J.-R. and M. Holt, 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

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.

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.

Dalenius, T. and J.L. Hodges Jr., 1959: Minimum variance stratification. Journal of the American Statistical Association, 54, 88-101.

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

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

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.

Murphy, A.H., 1991: Forecast verification: its complexity and dimensionality. Mon. Weather Rev., 119, 1590-1601.

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., 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.

Wald, A., 1943: Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 541, 426-482.

World Meteorological Organisation, 2012: Final report from the WMO Commission for Basic Systems Coordination Group for Forecast Verification; CBS/OPAG-DPFS/CG-FV. http://www.wmo.int/pages/prog/www/CBS-Reports/documents/CG-FV_Final-Report_May2012.pdf

MyWave

Proposal of metrics for developer and user focused verification of wave ensemble prediction systems

Reference: MyWave-D4.2b

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 : Andy Saulter, Chris Bunney, Paolo Pezzutto, Angela Pomaro	
MyWave version scope : All	
Approval Date : 02 Oct 2013	**Approver:** Andy Saulter
Dissemination level: Public	

DOCUMENT

VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:	Andy Saulter Chris Bunney	WP4 WP3	02 Oct 2013 02 Oct 2013
Distribution			
	Ø. Saetra (Project coordinator) A. Saulter (WP4) L. Cavaleri (WP3)		
	J.-R. Bidlot (WP4)		
	C. Bunney (WP3)		
	M. Gomez-Lahoz (WP3/WP4)		
	P. Pezzutto (WP3) A. Pomaro (WP3)		

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By

0.1	17/09/13	all	First draft of document	Andy Saulter	Chris Bunney;
					Marta Gomez-Lahoz;
					Paolo Pezzutto;
					Angela Pomaro

1.0	02/10/13	all	Document finalization	Andy Saulter	Chris Bunney

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABREVIATIONS

Baseline prediction	Prediction system used as a source of verification comparison
cdf	Cumulative distribution function
CRPS	Continuous Ranked Probability Distribution Score
DET	Detection Error Trade-off (plot)
DNV	Det Norske Veritas
EPS	Ensemble Prediction System
Hs	Significant wave height
kernel	Function used to estimate pdf of uncertainty data for a forecast
MAE	Mean Absolute Error
MCS	Marine Core Service
MDir	Mean wave direction
MST	Minimum Spanning Tree
Parameter space metrics	Verification measures that describe characteristics of prediction uncertainty in units of the parameter predicted
pdf	Probability distribution function
Probability space metrics	Verification measures that describe characteristics of prediction performance in terms of probability of a successful or unsuccessful prediction of an event
Q-Q	Quantile-quantile (plot)
Reference data	Data used to verify a prediction
RH	Rank Histogram
RMS	Root Mean Squared (value of parameter)
(R)MSE	(Root) Mean Squared Error
ROC	Relative Operating Characteristic (plot)
SAR	Synthetic Aperture Radar
SI	Scatter Index
SNRMSE	Symmetrically Normalised Root Mean Squared Error
T	(Generic) Wave period
Tp	Peak period of waves
Tz	Mean zero-upcrossing period of waves

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	MyWave-D4.2a	MyWave: Proposal of metrics for user focused verification of deterministic wave prediction systems	October 2013 / v1.1
DR 2	MyWave-WP4(UC)	MyWave: Categorisation scheme for MyWave users	March 2013 / v1.0

I INTRODUCTION

Tasks in MyWave WP4 will define operational verification methods that can be robustly applied within a wave element of a Marine Core Service (MCS). The purpose of this document is to propose metrics that provide model developers and (downstream) users with model performance or uncertainty data for (short-range) ensemble wave forecasts. Whilst there is significant overlap, the two communities are treated separately in this document to reflect that development of short-range high-resolution wave Ensemble Prediction Systems (wave-EPS) is an ongoing activity in the MyWave project (MyWave-WP3) and requires particular metrics in order to understand details of system configuration. On the other hand, the user community are identified as having a requirement for metrics that are particularly accessible (i.e. concisely presented and easily understood) and can be connected to application of wave forecast data. To that end, this document has been co-developed with colleagues in MyWave-WP3 in order to propose metrics critical to wave-EPS development (Section III), whilst further metrics are proposed as ‘user focused’ (Section IV) and will be reviewed as part of a process of user consultation described in Annex B.

This report accompanies a similar proposal for user focused metrics in report MyWave-D4.2a. In that report metrics were identified and categorised according to their **purpose**, i.e. the information that will be portrayed to the user. A similar approach is adopted here. Due to limitations in available observations a truly reliable analysis for waves is not available as a verifying reference, so the scope of this document is limited to verification applied to reference data comprising observations only. Furthermore we focus on metrics which can be ‘commonly’ derived against either in-situ or remote sensed sources for the most regularly sampled wave parameters. This is a pragmatic view which has been taken because high volumes of reference data are needed for ensemble verification and we will discuss metrics which will be applied in operational systems, where sample periods are likely to be limited to between a few months and a year.

The remainder of the document is set out as follows: in Section II the overall purpose of ensemble prediction and verification are described and some guiding principles that will influence the approach to metrics for MCS verification are set out; Section III presents metrics that will be important to the development of the wave-EPS systems in WP3; Section IV identifies further metrics that are likely to have resonance with users and will be evaluated through the user consultation process. The metrics are summarized in tables presented in Annex A. The user consultation process for verification metrics is outlined in Annex B.

II PRINCIPLES FOR ENSEMBLE PREDICTION AND MCS VERIFICATION

II.1 Purpose of Ensemble Prediction Systems (EPS)

In considering EPS verification it is useful to outline the purpose of an EPS and attributes of the system that should be tested. In numerical weather prediction (NWP) a deterministic forecast can be considered as a 'best guess' of future conditions. Significant effort will have been made to ensure that the starting estimate (analysis) of conditions is as accurate as possible and that the prognostic model uses the best available parameterizations of real-world physical processes. Nevertheless, both analysis and model physics will be subject to uncertainty. In theory, verification of the deterministic system allows determination of these uncertainties, however in practise sampling all the possible permutations of weather and ocean conditions in order to robustly describe uncertainty in any given situation is not possible.

Ensemble prediction aims to provide dynamically varying and accurate estimates of uncertainty on a forecast by forecast basis by sampling the uncertainty associated with weather system development and, if physics variations are introduced, model parameterization. The underpinning method involves running multiple instances of a forecast model from analyses where a level of variation has been permitted (and possibly physics; e.g. Bowler et al., 2008; Bonavita et al., 2008, 2010). The outcome is a set of discrete forecasts that should be sampled from the probability distribution function (pdf) of true conditions that could be realized a numbers of hours/days/weeks into the future.

Expected EPS properties which are to be verified should include:

- EPS members are a representative sample from the true pdf. One characteristic of this is that probability data for the occurrence of a given event in the EPS should have a direct relationship with the real-world probability of an event occurring (also known as forecast reliability; Murphy, 1993).
- Lower levels of uncertainty are predicted in the EPS when the evolution of conditions is stable and predictable (for example when a well established blocking high pressure system is in place) than in dynamic, unstable cases. This property links spread in the EPS to skill in a 'control' deterministic forecast.

- The EPS provides (at least) a qualitatively accurate guide as to whether conditions are more likely to be worse or better than predicted in the deterministic (control) case. If this is the case a forecast derived statistically using all EPS members should perform better than the control forecast.
- The EPS should have the ability to identify high impact situations with low probability at long-range and converge steadily toward these solutions at short-range in subsequent forecasts if real-world conditions develop in that way. This property is a function of the ability of the EPS base model to generate an extreme condition and reliability of the probability forecast at varying lead times.

In addition, the base model(s) used in the EPS will be subject to the same criteria as any deterministic model in terms of being able to accurately resolve the full range of climatological conditions without significant bias.

II.2 Comparison of approach to ensemble versus deterministic verification

The aim of ensemble verification is to test each of the properties described in Section II.1 and to demonstrate improvements to decision making enabled when using the EPS as an alternative to a deterministic forecast. In order to consider the basic approach to any of these tests Figure 1 presents a generalised form for an error pair used in verification where the reference is an observation. In the figure both prediction and reference values for a given parameter space (which for simplicity has been shown in a single dimension, but could be multi-dimensional or even circular) have uncertainties associated with them, which are shown in the form of pdfs. In the case of an EPS the forecast members should comprise a representative sample from the prediction uncertainty pdf. In the purely deterministic case the pdf is simplified to having a value of 1 at the predicted parameter value and zero elsewhere. The reference observation will be drawn from the joint probability distribution of the true condition plus an observation error.

Deterministic verification will either measure the (parameter space) distances between prediction and reference (i.e. the prediction errors) in order to quantify parameter uncertainty, or will measure the probability of a successful (or unsuccessful) forecast of dichotomous (yes/no) events based on predefined success criteria. Three further approaches are available in EPS verification that extend the deterministic case. In the first instance, where deterministic verification measures parameter space errors, EPS tests can be extended to assess whether the uncertainty distribution associated with each forecast is correlated with the occurrence of high or low errors, i.e. the spread-skill relationship. Secondly, parameter space error measurements can be applied to probabilities derived from the EPS (e.g. the Brier Score; Murphy, 1973), including measures of reliability. In the third case, where deterministic forecasts treat event prediction in a dichotomous (yes/no) manner, for an EPS the predicted event probability should vary between 0.0 and 1.0. This allows an extension to metrics that assess the ability of EPS members to identify events, and probability thresholds at which to make a deterministic decision about event occurrence. Besides, deterministic tests can be applied to deterministic forecasts derived statistically from the EPS member distribution (e.g. the ensemble mean value).

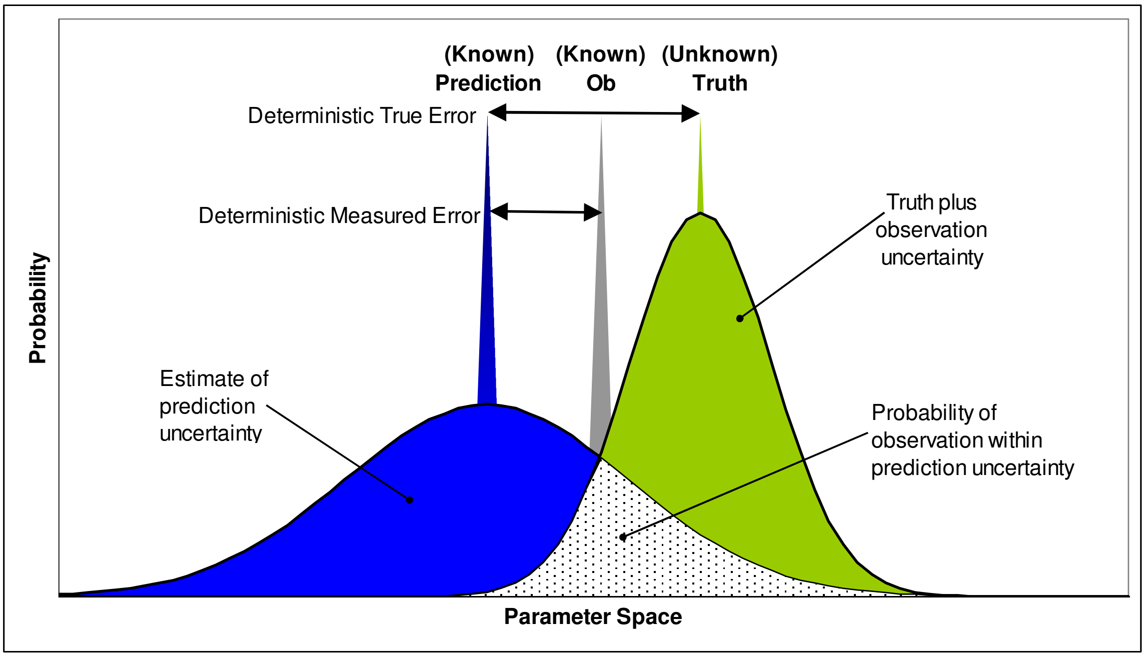
(Known)
Prediction
(Known)
Ob
(Unknown)
Truth
Deterministic True Error
Deterministic Measured Error
Truth plus observation uncertainty
Estimate of prediction uncertainty
Probability of observation within prediction uncertainty
Probability
Parameter Space

Figure 1. Schematic for parameter/probability space definition of prediction-observation error.

II.2.1 Prediction ‘dressing’

Potential applications of an EPS include those where the distribution of discrete ensemble members is used as a proxy for a continuous pdf. We may also wish to compare the EPS with deterministic forecasts in order to understand how the extra information in the EPS might improve decision making, and in these cases it should be considered that in practical usage deterministic forecasts will be treated 'with a pinch of salt' and some form of uncertainty will be implicitly built into decision making using deterministic data. Figure 1 acknowledges that observations used as a reference will be subject to a level of uncertainty relative to the 'true' sea-state. Each of these factors can be taken into account by application of suitable 'dressing' of the verified forecasts. However, the schemes used should be mindful of the fact that the aim is to verify the EPS forecasts and not a post-processed version of the data.

As a result it is proposed that dressing within the context of MyWave verification should be approached as simply as possible. Dressing applied to the EPS should concentrate principally on assessing the effects of observation errors on occurrence of EPS-observation outliers in the tail of the estimated pdf (following Saetra et al., 2004). Baseline probabilistic forecasts constructed from deterministic systems should be based on past verification data (e.g. Flowerdew et al., 2010). In both cases the kernel function used to dress the data can be assumed to take a relatively simple form, such as a Gaussian.

II.2.2 Application of observation errors

Two types of observed errors can be considered, namely 'representation errors' and 'instrument errors'. Representation errors, which relate to the scales over which waves are sampled by the observing system or represented in a model, can be mitigated in a verification scheme by standardising the scales over which observations are aggregated as much as possible. Achieving consistency in this respect is desirable as it means that the role of representation errors in the verification will only vary with the model used. Previous work has used an estimated model scale to determine aggregation schemes for observations (e.g. Bidlot and Holt, 2006; Bidlot et al., 2007). However, for a MCS verification system which potentially compares different models, and bearing in mind the common perception that in-situ observations are the de-facto standard for true sea-state, it can be argued that referencing against a 15-30 minute in-situ sample enables more consistent and user focused intercomparison. Based on propagation speeds for wave energy, this length of sample at a fixed point is equivalent to an approximate 8-25 km area sample for waves (in deep water) of period 5-16 seconds and, for example, would be represented by an aggregation over 2-4 data points from satellite altimeter soundings at 1Hz frequency.

Instrument errors comprise systematic and random components. Previous studies have demonstrated that these errors can be estimated via assessment of the observations over a long time period (e.g. through triple collocation studies; Janssen et al., 2007). An assessment of in-situ and satellite altimeter errors and their subsequent application to verification metrics is part of ongoing work within MyWave-WP4 and for ensemble metrics will follow the philosophy employed by Saetra et al. (2004). The basic principle is to assume that the EPS samples uncertainty in the development of the true conditions and that instrument errors are independent. In this case the prediction versus reference pdf should be properly represented when instrument uncertainty estimates are convolved with the ensemble member distribution in order to generate a forecast pdf.

II.2.3 Baseline forecasts

In certain cases, particularly when using scoring metrics, it may be important to contextualise the verification by referencing against another 'baseline' prediction system. The eligible baselines are highlighted:

- The verifying reference climate will provide a perfectly reliable and resolved prediction of the reference dataset, but will not vary dynamically from forecast to forecast. These data are incorporated in uncertainty components of scores such as the Brier Score (Murphy, 1973) and Continuous Ranked Probability Score (Hersbach, 2000).

- The EPS control member plus an associated deterministic uncertainty measure (e.g. errors from past verification) will provide a dynamic forecast but with an estimated (and not physically representative) degree of spread. This comparison may be particularly useful in situations where forecast uncertainty might be expected to be particularly constrained, for example at very short lead times or when depth or topography limits the wave climate that can be achieved.

- Long term model climatology will provide forecasts that should resolve climate in a similar fashion to the EPS, but may be neither a reliable or well resolved estimate of the verifying reference climatology. These forecasts will not change dynamically, but could be assumed to be a best available predictor if no forecast system were available.

II.3 Practical considerations for wave-EPS verification

Within both the scope of wave-EPS development in the MyWave project and potential requirements for regularly updated operational verification a number of practical considerations will constrain the metrics and parameters that can be verified.

A major constraint is that probabilistic forecast metrics are 'data hungry', particularly in instances where reliability, spread-skill relationships or probability thresholds are being evaluated. In these instances both an adequate range of reference conditions and EPS probabilities of occurrence need to be (independently) sampled. MyWave-D4.2a discusses sampling requirements for deterministic metrics and suggests that the minimum sample period for evaluation of forecasts against in-situ or satellite observed baselines in regional seas should be 3 months. Based on a relationship derived by Candille and Talagrand (2004), in order to estimate reliability to a precision of 10% based on an ensemble of 20 members a sample size of approximately 14000 events is required. On this basis a recommended sample period for wave-EPS verification is at least 3-6 months and is applicable only for summary statistics from univariate analysis of the most regularly sampled parameters. The sample period may need to be extended further for testing multiple variables together and assessing reliability. Parameters considered as regularly sampled are:

- Significant wave height (from either in-situ or satellite altimeter data)
- Wave period (where a sufficient number of in-situ platforms exist)
- Wave direction (where a sufficient number of in-situ platforms exist)
- Wind speed (from either in-situ, satellite altimeter or scatterometer data)
- Wind direction (from either in-situ or scatterometer data).

Using the reference data in a manner that is common to both in-situ and satellite remote sensed data requires that the sample of events verified comprise instantaneous 'snapshots' of given parameters at specific locations/times. Verification that uses these data cannot assume or make use of any spatial or temporal linkage between events. In reality, if the sampling rate is high, such links will be present and it may then become important to ensure that the sample used is not aliased by particular sub-collections of data within the sample (for example if in-situ data within a region are clustered in a particular area). In MyWave-D4.2a it has been proposed that the use of a block bootstrap approach (Carlstein, 1986; Kunsch,

1989) be explored in order to ensure that a series of independent data blocks are used in construction of metrics whilst retaining the maximum number of observations within the verifying sample. A similar approach could be explored for wave-EPS metrics (e.g. Candille et al., 2007), although this leads to extra complexity in both processing and presentation of the verification data.

A further consideration in presentation of verification will be that, as for the deterministic case, some conditional sampling effects on the metrics might be expected. Use of a suitable resampling scheme (e.g. block bootstrap) to assess some aspects of sampling effects and retaining a rolling record of certain metrics in order to assess long term variability in the data are suggested options to help understand and explain these effects to users.

II.4 Principles for Marine Core Service verification

Metrics in this report are expected to be used both to assess and demonstrate performance of wave-EPS data during systems development and to provide performance data for users within an operational framework (as part of a MCS). The key attributes of a wave-EPS that the metrics should verify are discussed in subsection II.1, but it is proposed that MCS verification should be mindful of some further guiding principles regarding usage and portrayal of verification data:

- The system may need to be considerate of the fact that service users might wish to apply verification data within downstream services or decision making processes.
- Metrics should be regularly updated to reflect recent system performance. For example in the MyOcean service metrics are updated every 3 months and are presented in a rolling archive of up to 1 year of data (Alistair Sellar, *pers. comm.*).
- The system should enable rapid discovery of metrics that allow downstream users to easily understand performance of the prediction system relevant to their particular use of the MCS data.
- The metrics should be accessible to non-scientific users; for example, if the metrics provided cannot be explained with a few sentences of text, they are probably not fit for purpose.
- Metrics comparing prediction system performance against a baseline prediction should be meaningful in terms of user decision making.

Adopting these principles means that there is a need to clearly associate given metrics with an application that users can recognise, and also to ensure that verification data which can practicably be made available within an operational verification scheme covers as many key user applications as possible. To this end the metrics presented in this document will be classified according to the purpose that each aims to fulfil. Clearly defining what each metric does is important to MCS application since, in general, it is expected that users are unlikely to wish to review large numbers of metrics and will instead want to quickly discover those key pieces of verification data that meet a specific need.

In this document four overarching purpose categories for EPS verification are identified according to the aspect of model performance being tested:

- Climatology tests (Annex A, Table EC) determine the ability of the prediction system to replicate the reference climate, for example describing sharpness and bias of the predictions. These tests ignore any time-referencing in the sample pairs. The outcomes may be used to determine systematic errors and specific process representation issues and, in the context of an EPS, can be used to assess the underpinning model and any statistically derived predictor (e.g. the ensemble mean).
- Measures of prediction uncertainty in parameter space (Annex A, Table EM) estimate accuracy from the sample of prediction-reference errors in the deterministic case. These metrics enable the errors to be viewed in context against background conditions or in prediction system intercomparison. In the EPS context these are extended to assess the relationship between EPS spread and deterministic forecast uncertainty and to summarize probability errors and reliability.
- Measures of (dichotomous) prediction uncertainty in probability space (Annex A, Table EP) describe the ability of a prediction system to successfully identify given reference conditions. These data can be used to evaluate the long term benefits of using the predictions (i.e. whether more gains than losses will be made through basing decisions on prediction data). For an EPS these are extended to assess use of EPS event probability prediction data as the decision making system.
- Assessment of performance in forecasting extreme conditions (Annex A, Table EX) analyse performance of the model specifically at the tail(s) of the distribution of conditions. The tests described are intended to be robust when working with limited data samples.

Metrics falling into each purpose category are identified within the following subsections. Tabular summaries of the full set of proposed metrics by purpose category are given in Annex A.

III PROPOSED METRICS FOR WAVE-EPS DEVELOPMENT

Metrics proposed in this section comprise a set of core tests that are anticipated to be required in order to validate a successful implementation of short-range wave-EPS systems with MyWave-WP3.

III.1 Deterministic Metrics

Deterministic metrics test properties of either individual members of the EPS, or products derived statistically from the distribution of EPS members that can be used in a deterministic fashion (e.g. the ensemble mean). The metrics enable intercomparison between members, against other deterministic forecast systems, or against naïve predictors such as chance or climatological mean. A range of deterministic metrics are discussed in MyWave-D4.2a, from which the metrics considered most pertinent to developing the wave-EPS systems are proposed below.

In particular it will be useful to understand if inequalities in predictive skill exist between members, for example if there are significant deviations in performance of individual members compared to the control (for example as a result of the inclusion of lagged forecast members in the Met Office wave-EPS in MyWave-WP3). In principle individual members of the EPS should adequately replicate the reference climate, whilst forecasts derived statistically from the ensemble (e.g. the ensemble mean) may be limited in this sense they are not direct simulations of the physical environment. This may be particularly true for replication of conditions within the tail of the distribution. Intercomparison of the deterministic metrics enables this, although in an ensemble consisting of more than a few members graphical presentation will allow the most accessible and concise view of the data.

Test EC1: Reproduction of general features of the reference climate

The most concise metrics are based on comparing moments of the event sample distributions and should include higher moments of the distribution relating to skewness and kurtosis since many parameters being tested (e.g. significant wave height, wind speed) cannot be assumed to be normally distributed.

Proposed metrics (in combination):

- Parameter mean, $\mathrm{E}[x]=\frac{\sum x}{n}$ (for variable x with sample size n); differentials in reference and predicted means measure bias
- Parameter root mean squared (RMS) value, $\mathrm{RMS}[x]=\sqrt{\frac{\sum x^2}{n}}$
- Parameter standard deviation $\sigma=\sqrt{\mathrm{Var}[x]}=\sqrt{\frac{\sum (x-\mathrm{E}[x])^2}{n}}$
- Parameter skewness $\gamma=\mathrm{E}\left[\left(\frac{x-\mathrm{E}[x]}{\sigma}\right)^3\right]$
- Parameter kurtosis, $\beta=\mathrm{E}\left[\left(\frac{x-\mathrm{E}[x]}{\sigma}\right)^4\right]$, or kurtosis exceedence from the normal distribution value, i.e. β - 3

Test EC2: Reproduction of details of the reference climate

Distribution comparisons can be used to provide more detail in representation of the reference climate and highlight sub-ranges of conditions which are particularly well or poorly replicated. Quantile-quantile (Q-Q) plots are recommended for development as these provide a useful visualization for the distribution tails.

Proposed metrics:

- Q-Q plot; for parameters with long distribution tails (e.g. significant wave height) split over two levels to resolve body and tail of distribution

Test EM1: Quantify the scale of errors

In parameter space, Root Mean Squared Error ($RMSE$) and Mean Absolute Error (MAE) are the most recognised metrics for overall error description. $RMSE$, which is a composition of bias and a measure of error scatter, is a particularly popular metric, but has been demonstrated to have drawbacks when comparing data with similar levels of performance (Mentaschi et al., 2013). As a result it is recommended that $RMSE$ is presented alongside a breakdown of contributions to the metric as described in Test M1a. Mentaschi et al. (2013) also discuss use of a corrected normalised indicator following Hanna and Heinold (1985), which mitigates issues with $RMSE$ by symmetrically normalising the squared error data using both prediction and reference values.

Proposed metrics:

- Mean Absolute Error, $MAE = \frac{\sum |EP|}{n}$, where EP denotes the sample of errors for prediction (M) and reference (R), ($EP_i = M_i - R_i$)
- Root Mean Squared Error (as for parameter RMS with EP as the input variable)
- Hanna and Heinold (1985) symmetrically normalised $RMSE$; $SNRMSE = \sqrt{\frac{\sum EP_i^2}{\sum M_i R_i}}$

Test EM1a: Assess effects of prediction 'sharpness and reliability' on RMSE

Reviewing the contribution to $RMSE$ of prediction variability, correlation or bias is expected to be useful to model developers studying the overall effects of system changes. Mean Square Error (MSE) comprises bias and error variance contributions as

$$MSE = \mathrm{Var}[EP] + \mathrm{E}[EP]^2 ,$$

where error variance further breaks down as:

$$\mathrm{Var}[EP] = \mathrm{Var}[R] + \mathrm{Var}[M] - 2\,\mathrm{Cov}[M, R]$$

MSE can be normalised by Var[R] (to give a skill score relative to a naïve predictor based on the reference mean). The normalised variance component is a form of (squared) Scatter Index (SI, which has also been defined in other forms by Bidlot et al., 1997; Ardhuin et al., 2007; Filipot and Ardhuin, 2012). Breaking down the $SI_{RVar}{}^2$ used here gives:

$$SI_{RVar}{}^2 = 1.0 + \frac{\mathrm{Var}[M]}{\mathrm{Var}[R]} - 2\frac{\mathrm{Cov}[M, R]}{\mathrm{Var}[R]}$$

in which the third term can be re-written in terms of correlation and variance using:

$$\frac{\mathrm{Cov}[M,R]}{\mathrm{Var}[R]} = \mathrm{Corr}[M,R]\sqrt{\frac{\mathrm{Var}[M]}{\mathrm{Var}[R]}}$$

The normalised prediction variance and correlation can, respectively, be viewed as measures of the prediction systems' sharpness (i.e. how much the prediction attempts to replicate the reference 'signal') and reliability (i.e. whether the prediction is able to track the reference as it transitions through the range of conditions). In an ideal situation the normalised MSE will be reduced when both the normalised prediction variance and the correlation tend to 1.0 (so that ${SI_{RVar}}^2$ tends to 0.0), and when the bias part tends to 0.0. However the relationship for ${SI_{RVar}}^2$ is minimised when normalised prediction standard deviation is equal to the correlation value and therefore MSE will favour prediction systems with lower variance as correlation reduces. Mentaschi et al. (2013) also demonstrate dependence between SI and bias, such that SI is reduced in cases where the prediction has a negative bias. It can be argued that for wave prediction neither a reduction in forecast sharpness or a tendency to under-predict are desirable qualities, and so the MSE breakdown as described should help to indicate if reduced $RMSE$ scores have resulted from either of these effects. When many predictions are being compared the Taylor plot (Taylor, 2001) provides a useful visualization of the ${SI_{RVar}}^2$ breakdown.

Proposed metrics (in combination):

- MSE normalised by reference variance
- Bias normalised by reference variance
- (Squared) Scatter Index, ${SI_{RVar}}^2$
- Pearson Correlation
- Standard deviation of prediction normalised by reference standard deviation
- Taylor plot

Test EP1: Quantify deterministic ability to predict event x

This test is expected to be particularly applicable for comparing deterministic performance of ensemble control and mean, although it could also be used to detect performance differentials between members. For model development the main issue is selecting a suitable series of events against which the tests can be conducted. The basis for this test is a 'contingency table' for a dichotomous (yes/no) forecast as presented below:

	Event observed	Event not observed
Event predicted	*Hit*	*False Alarm*
Event not predicted	*Miss*	*Correct Rejection*

Within MyWave-D4.2a it was proposed that the MCS verification scheme would also publish a small set of critical and accessible parameters. Initially these are identified as:

$FractionCorrect = \dfrac{Hits + CorrectRejections}{SampleSize}$, which quantifies the chance that predictions successfully identify both events and non-events.

$SuccessRatio = \dfrac{Hits}{Hits + FalseAlarms}$, which quantifies the chance that an event will occur if predicted.

$FalseAlarmRatio = 1 - SuccessRatio$, which quantifies the chance that an event will not occur if predicted.

$MissRatio = \dfrac{Misses}{Misses + CorrectRejections}$, which quantifies the chance of an event occurring if not predicted.

Proposed metrics:

- Contingency table for event
- Percentage scores for: *Fraction Correct*, *Success Ratio*, *False Alarm Ratio* and *Miss Ratio*

III.2 Testing properties of ensemble spread

The metrics in this subsection test the ability of the ensemble spread to dynamically predict uncertainty associated with a deterministic prediction of the reference conditions. In general it is expected that the spread measure used is associated with the choice of deterministic prediction, i.e. relative to either the control member or ensemble mean.

Test EC3: Quantify forecast to forecast variability in EPS spread

This climatology test evaluates mean and variance of EPS spread from forecast to forecast and provides a first check that the EPS spread is a useful dynamic quantity. Since the test relies only on EPS data, gridded visualizations of the metrics can be generated. The baseline reference in this case would be a spread measure applied to a long term model climatology of the parameter being tested.

Proposed metric(s):

- Mean spread (site/area specific or mapped)
- Standard deviation of spread (site/area specific or mapped)
- Spread histogram

Test EM2: Quantify probability that ensemble spread captures variability of the reference

A simple and accessible method of testing whether ensemble spread is sufficient to capture variability associated with the reference is to use a bounding box metric (Weiseheimer et al., 2005) in which the EPS is successful if the reference data are captured within the range covered by EPS members. The approach can be extended to multi-variate cases. The metric may be sensitive to systematic biases in the EPS however. This metric allows comparison of the EPS against either long term model climatology (as test for systematic issues in predicting extremes) or a dressed deterministic forecast (where the bounding box is defined by a window around the deterministic forecast value) as a baseline prediction.

Proposed metric:

- Bounding box

Test EM3: Describe characteristics of overspread or underspread in the EPS

More detailed assessments of EPS spread characteristics relative to the reference can be made by assessing where the reference value falls relative to a list of EPS members ranked by parameter value. In the univariate case the data are graphically presented using a Rank Histogram (RH; Talagrand and Vautard, 1996; Hammill and Collucci, 1997) and can be extended to multi-variate cases, for example using distances derived with a Minimum Spanning Tree method (Wilks, 2004; Gombos et al., 2007).

Proposed metric(s):

- Rank Histogram (extension to MST in multi-variate cases)

Test EM4: Describe the relationship between EPS spread and deterministic forecast errors

If tests EM2 and EM3 provide sensible results then a comparison between the EPS forecast spread and errors for a deterministic forecast (e.g. control member, ensemble mean) will illustrate whether changes in the ensemble spread successfully discriminate levels of deterministic forecast uncertainty. The measures that can be used for both spread and error are somewhat flexible, for example Scherrer et al. (2004) compare RMS values of both EPS spread and deterministic error whilst Saetra and Bidlot (2004) compare absolute errors with EPS inter-quartile range.

Proposed metric(s):

- Spread-Skill scatterplot for EPS spread versus deterministic error
- Relationship fit to Spread-Skill data

III.3 Testing properties of probabilistic forecasts

The final set of development tests assess the viability of using quantitative probabilities derived from the EPS as a prediction of the reference state. Two forms of testing are identified which a) evaluate errors in the EPS probability forecast of the reference data (and are effectively measures in parameter space), b) assess the use of EPS probabilities as a dichotomous (yes/no) forecast method.

Prediction baselines for these tests can be long term model or reference climatology, or a dressed deterministic forecast. For model development the main issue is selecting a suitable set of events against which the dichotomous forecast tests can be conducted.

III.3.1 Testing that probabilities in the reference data are replicated

Test EM5: Summarize performance of probabilistic forecasts in parameter space

The Continuous Ranked Probability Score ($CRPS$, Hersbach, 2000) provides a summary measure of probability forecast errors in parameter space that can be viewed as an extension to the deterministic Mean Absolute Error. The $CRPS$ is constructed for an individual case using:

$$CRPS(P, x_r) = \int_{-\infty}^{\infty} [P(x) - P_r(x)]^2 \, dx$$

Where $P(x)$ and $P_r(x)$ are cumulative distributions for the prediction and reference and:

$P(x) = \int_{-\infty}^{x} \rho(y) dy$, for the predicted pdf $\rho(x)$

$P_r(x) = \mathrm{H}(x - x_r)$

where H is the Heaviside function,

$$H(y) = \begin{cases} 1 \text{ for } y \geq 0 \\ 0 \text{ for } y < 0 \end{cases}$$

The metric will reward an EPS that is accurate and limits its spread as much as possible. The data can be presented either as a distribution of $CRPS$ scores or as an overall mean. Hersbach (2000) also describes a decomposition of the score into reliability, uncertainty and resolution components. The mean of the reliability estimates the degree to which the cumulative distribution function(s) (cdf) from the ensemble forecast reflects the reference cdf. The uncertainty score is the CRPS for the reference climatology and resolution quantifies the level to which the EPS improves on the reference climatology as a naïve forecast.

Proposed metrics:

- Mean Continuous Ranked Probability Score ($CRPS$)
- $CRPS$ reliability, resolution and uncertainty
- Distribution of $CRPS$ scores

III.3.2 Testing application of probability data to dichotomous event forecasting

Test EP2: Compare the number of EPS members predicting event x with rates of forecast success

This test allows the model developer to assess both the rates at which the EPS members will identify given conditions, and the associated outcomes regarding occurrence or non-occurrence of the event in the reference data. The method proposed is to extend the standard deterministic contingency tables such that the table rows correspond to number of members predicting the event, i.e.:

Number of members predicting event	Event Observed	Event Not Observed
N		
N-1		
…	…	…
1		
0		

Proposed metric(s):

- Extended contingency table

Test EP3: Summarize probability forecast errors for event x

The Brier Score (BS; Murphy, 1973) is the mean squared error of the probability forecast:

$$BS = \mathrm{E}\left[(\rho_e - r_e)^2\right],$$

where ρ_e is the predicted probability of event e and r_e is set to either 1 or 0 dependent upon whether the event occurred or not. Murphy (1973) also demonstrates a decomposition into reliability, uncertainty and resolution components, where reliability compares forecast probabilities to observed relative frequencies, uncertainty describes the variance of reference value frequency in the sample and resolution estimates the ability of the EPS to issue reliable forecasts with very high or low probability values.

Proposed metric:

- Brier Score
- Reliability, resolution and uncertainty decomposition

Test EP4: Describe probability forecast reliability for event x

Where enough data are available the Reliability diagram (Wilks, 1995) provides a direct visual comparison between EPS forecast probabilities and associated rates of occurrence in the reference sample.

Proposed metric:

- Reliability diagram

IV USER FOCUSED METRICS

User focused metrics should provide an accessible summary of EPS performance characteristics and present verification that helps users understand prediction uncertainty in terms of its application to decision making. The baseline prediction for a number of these tests is ideally a deterministic forecast (with an appropriate form of dressing applied) since this is most likely to be the available alternative forecast system.

For deterministic elements derived from the EPS (control, mean) any of the user focused tests described in MyWave-D4.2a are appropriate. In particular it is suggested that users may want to compare the distributions of control versus ensemble mean errors (Tests M2, M3 and R1) and replication of extreme events (Tests X1, X2). In addition, from the metrics described in the Section II for EPS development, the following tests are proposed as suitable for user focused verification:

- EC1: Intercomparison of climate metrics for individual ensemble members plus mean
- EM1/1a: Intercomparison of RMSE and breakdown for individual members plus mean
- EM2: Quantify the probability that EPS spread captures variability of the reference data
- EM4: Describe the relationship between EPS spread and deterministic forecast errors
- EM5: Summarize performance of the probabilistic forecast in parameter space
- EP1: Quantify deterministic ability to predict event x
- EP2: Compare the number of EPS members predicting event x with rates of forecast success
- EP3: Summarize probability forecast errors for event x
- EP4: Describe probability forecast reliability for event x

IV.1 Further metrics for probability forecasts

Test EM6: Quantify the ability of an EPS probability forecast to identify a reference event within given bounds

Describing the probability with which the EPS predicts reference conditions (within predefined bounds) provides a very accessible metric which also has a direct analogue to testing that can be carried out on a deterministic model (Test P1 in MyWave-D4.2a). The basis for the metric is the Wilson Score (WS; Wilson et al., 1999) which calculates the probability from the EPS forecast for the reference event within predefined bounds, and is described for a given forecast as:

$$WS = \int_{x_r - \Delta x_r}^{x_r + \Delta x_r} \rho(x) dx .$$

The metric is influenced by the choice of error bounds applied to the reference, and therefore allows the metric to be presented in such a way that the user can compare the range expected for reference outcomes against levels of probability regularly predicted by the EPS.

Proposed metric:

- Mean Wilson Score
- Distribution of WS values

Test EP5: Quantify effect of probability threshold for forecasting event x

A guide to the effects of varying the probability threshold on decision making can be provided visually to the user via Relative Operating Characteristic curves (ROC, e.g. Mason, 1982; Buizza and Palmer, 1998) which compare *Probability of Detection* (chance of correctly forecasting an event) against *False Alarm Rate* (chance of forecasting an event that did not occur), or alternatively a Detection Error Trade-off curve (*Miss Ratio* versus *False Alarm Rate*; Martin et al., 1997) for cases where users are more interested in ensuring that an event is not missed.

Proposed metric:

- Relative Operating Characteristic plot

- Detection Error Trade-off plot

Test EP6: Test long term value of using probability forecasts in decision making

Some users may wish to apply their own cost-loss models in order to verify the impact that EPS based decision making will have on their operations, for which the extended contingency table (Test EP2) will be relevant. It is also suggested that a generalised comparison of cost-loss benefits, using a simple cost-loss assessment in which a predicted event is associated with a cost (the same value is taken for a false alarm or a hit) and any miss is associated with a loss, will be a useful presentation in MCS verification. This cost-loss assumption allows an Economic Value score to be generated against a varying cost-loss ratio (C/L in the range 0 to 1) since the cost of the prediction system will be:

$$EV = C.(Hits + FalseAlarms) + L.Misses$$

Relative scores can be generated by referencing against a baseline prediction. Carrasco et al. (2013) discuss application of a relative score, following Richardson (2000), that is constructed from costs associated with a situation in which no forecasts are available (in the case where action is never taken the cost will be $EV_c = L(Hits + Misses)$) and a perfect forecast (cost is $EV_{perfect} = C(Hits + Misses)$), so that Relative Economic Value:

$$REV = \frac{EV_c - EV_{EPS}}{EV_c - EV_{perfect}}.$$

REV values can be plotted in comparison to one another for various forecast strategies, e.g. EPS versus deterministic and for various threshold criteria.

Proposed metric:

- Relative Economic Value plot

IV.2 Application to extreme cases

For extreme cases the main issue for generating the metrics is sample size. Atger (2004) discusses the use of a fitted relationship to the ROC curve in order to then estimate forecast reliability. However, within an operational scheme presented to users adopting more simple metrics that trace recent events is proposed as a more accessible approach. It is suggested that these metrics are applied to events above the 95th percentile of the reference distribution.

Test EX1: Quantify the number of extreme events that were predicted within bounds of the EPS forecast

This test uses the Bounding Box principle to establish how often extreme conditions are successfully identified within the scope of the EPS members. This should provide users with a simple first estimate of whether the EPS suffers from any form of low bias.

Proposed metric:

- Bounding Box

Test EX2: Quantify deterministic errors between extreme events and maximum / mean / control ensemble members

The purpose of this test is to quantify systematic errors between reference and various deterministic forecast indicators obtained from the EPS. In particular these tests should highlight variations between forecasts representing the centre of the EPS sample and the upper bounding member.

Proposed metrics:

- Error distribution
- Bias
- MAE

Test EX3: Quantify the number of EPS members indicating event

This test reviews the number of EPS members that indicated reference extreme events, based on preset criteria. The criteria can be quantitative in parameter space (e.g. using a significant wave height threshold for high wave events) or in terms of a predefined climatology (e.g. in the manner used in ECMWF’s Extreme Forecast Index; Lalaurette, 2003; Petroliagis and Pinson, 2012). Presenting several criteria should enable users to make sensible judgements on how to employ the EPS to ensure that the risk of missing an extreme event is minimised, but in this case the converse metric that presents the trade off between Miss Ratio and False Alarm Rate should be provided.

Proposed metric:

- Distribution of member numbers achieving event indication criteria for reference events
- Detection Error Trade-off plot for set criteria

V SUMMARY AND NEXT STEPS

This document proposes a set of metrics that the MyWave project will test in verification of wave-EPS data, both during system development and for the purpose of communicating the main aspects of system performance to users under operational conditions. In order to identify the relevance of the metrics to wave-EPS development or user focused task, the purpose of each metric presented has been defined. These definitions are considered particularly important to MCS portrayal of the verification since they should enable rapid discovery of relevant metrics by different user types. A summary of the metrics is presented in the tables in Annex A.

A number of technical considerations for EPS verification have been discussed at a high level, and will require consideration and a more detailed approach to be adopted within MyWave-WP3. In particular a limited approach to dressing the EPS members has been suggested in order to ensure that the raw ensemble is verified rather than a post-processed version of the data. The approach follows Saetra et al. (2004) and concentrates on applying effects of observation errors to the discretised sample of conditions forecast by ensemble members. This will be discussed in more detail in later reports from WP3 and WP4.

A major constraint on the metrics and sample period they represent is the availability of observed reference data. In order to calculate statistics with a reasonable level of accuracy it is expected that at least 3 month samples of the most common observed parameters will be required for regional sea areas, and that this will need to be increased to 6 month or 12 month periods where multi-variate, reliability or extreme data are tested.

Assumptions about both the user requirements for certain metrics and technical feasibility of implementation and portrayal will require testing. This will be carried out through the process of user consultation and evaluation of proposed metrics as discussed in Annex B of this document.

VI REFERENCES

Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. Q.J.R. Meteorol. Soc., 130: 627–646. doi: 10.1256/qj.03.23

Ardhuin, F., L. Bertotti, J.-R. Bidlot, L. Cavaleri, V. Filipetto, J.-M. Lefevre and P. Wittmann, 2007: Comparison of wind and wave measurements and models in the Western Mediterranean Sea, Ocean Eng. 34(3-4), 526-541

Bidlot, J.-R. and M. Holt, 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

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 Workshop on Wave Hindcasting and Forecasting and Coastal Hazard Symposium, North Shore, Oahu, Hawaii, November 11-16, 2007.

Bonavita, M., L. Torrisi and F. Marcucci, 2010: Ensemble data assimilation with the CNMCA regional forecasting system. Q. J. R. Meteorol. Soc., 136, 132-145

Bonavita, M., L. Torrisi and F. Marcucci, 2008: The ensemble Kalman filter in an operational regional NWP system: Preliminary results with real observations. Q. J. R. Meteorol. Soc., 134, 1733-1744

Bowler, N. E., A. Arribas, K.R. Mylne, R.B. Robertson and S.E. Beare, 2008: The MOGREPS short-range ensemble prediction system. Q. J. R. Meteorol. Soc., 134, 703– 722

Buizza, R. and T.N. Palmer, 1998: Impact of ensemble size on ensemble prediction. Mon. Weather Rev., 126, 2503–2518

Candille, G., and O. Talagrand, 2004: Impact of observational errors on the validation of ensemble prediction systems. Ensembles Workshop, Exeter, United Kingdom.

Candille, G., C. Côté, P. L. Houtekamer, and G. Pellerin, 2007: Verification of an Ensemble Prediction System against Observations. Mon. Wea. Rev., 135, 2688–2699.

Carrasco A., Ø. Sætra and J.-R. Bidlot, 2013: Cost-loss analysis of calm weather windows. Journal of Operational Oceanography, Vol 6 No 1,17-22.

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.

Filipot, J.F. and F. Ardhuin, 2012: A unified spectral parameterization for wave breaking: from the deep ocean to the surf zone. J. Geophys. Res., 117, C00J08, doi:10.1029/2011JC007784

Flowerdew, J., K. Horsburgh, C. Wilson, and K. Mylne, 2010: Development and evaluation of an ensemble forecasting system for coastal storm surges. Q.J.R. Meteorol. Soc., 136: 1444–1456. doi: 10.1002/qj.648

Gombos, D., J.A. Hansen, J. Du and J. McQueen, 2007: Theory and applications of the Minimum Spanning Tree Rank Histogram. Mon. Wea. Rev., 135, 1490-1505.

Hamill, T., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. Mon. Wea. Rev.,125, 1312–1327.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Wea. Forecasting, 15, 559–570.

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.

Lalaurette F. 2003. Early detection of abnormal weather using a probabilistic Extreme Forecast Index. Q. J. R. Meteorol. Soc. 129: 3037–3057.

Martin, A. F., G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, 1997: The DET Curve in Assessment of Detection Task Performance. Proc. Eurospeech '97, Rhodes, Greece, September 1997, Vol. 4, pp. 1899–1903.

Mason, I., 1982: A model for assessment of weather forecasts. Aust. Meteorol. Mag., 30, 291–303.

L. Mentaschi, 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

Murphy, A.H., 1973: A new vector partition of the probability score. J. Appl. Meteor.,12, 595–600.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Wea. Forecasting, 8, 291-293.

Petroliagis, T. I. and P. Pinson 2012: Early warnings of extreme winds using the ECMWF Extreme Forecast Index. Met. Apps. doi: 10.1002/met.1339

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. Quart. J. Royal Met. Soc., 126, 649-667.

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, and D.S. Richardson, 2004: Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability. Mon. Wea. Rev., 132, 1487–1501.

Scherrer, S.C., C. Appenzeller, P. Eckert and D. Cattani, 2004: Analysis of the Spread–Skill Relations Using the ECMWF Ensemble Prediction System over Europe. Weather and Forecasting, 19, 552-565.

Talagrand, O., and R. Vautard, 1997: Evaluation of probabilistic prediction systems. Proc. ECMWF Workshop on Predictability, Reading, United Kingdom, ECMWF, 1–25.

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.

Weisheimer, A., L.A. Smith. and K. Judd, 2005: A new view of seasonal forecast skill: bounding boxes from the DEMETER ensemble forecasts. Tellus A, 57: 265–279. doi: 10.1111/j.1600-0870.2005.00106.x

Wilks, D.S., 1995: Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press, 467 pp.

Wilks, D.S., 2005: The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. Mon. Wea. Rev., 132, 1329-1340.

Wilson, L.J., W.R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. Mon. Wea. Rev., 127, 956-970.

VII ANNEX A: TABLES OF PROPOSED METRICS

Tables in this section summarise metrics that are proposed to be tested within development of new wave-EPS in MyWave WP3 and in the MyWave user consultation on verification as part of WP4. Deterministic products from the EPS may also be tested using the metrics described in MyWave-D4.2a.

Table EC: Climatological Tests

Purpose	Proposed Metric(s)	Feature of EPS verified	Target Audience
EC1: Test that general features of the reference climate are reproduced	Combined comparison of parameter: Mean; RMS; Standard Deviation; Skewness; Kurtosis Exceedence	EPS members Ensemble mean	Developer + User
EC2: Test that details of the reference climate are reproduced	Q-Q plot	Control member Ensemble mean	Developer
EC3: Quantify forecast to forecast variability in EPS spread	(Spread) Mean; Standard Deviation; Histogram	EPS spread	Developer

Table EM: Measures of prediction uncertainty in parameter space

Purpose	Proposed Metric(s)	Feature of EPS verified	Target Audience
EM1: Quantify the scale of errors	MAE RMSE SNRMSE Bias	EPS members Ensemble mean	Developer + User
EM1a: Assess effects of prediction 'sharpness and reliability' on RMSE	Combined: Normalised MSE; Normalised Bias; (Squared) SI; Pearson Correlation; Normalised Standard Deviation	EPS members Ensemble mean	Developer + User
EM2: Quantify probability that ensemble spread captures variability of the reference	Bounding Box	EPS spread	Developer + User
EM3: Describe characteristics of overspread or underspread in the EPS	Rank Histogram (univariate and multivariate) Minimum Spanning Tree	EPS spread	Developer
EM4: Describe the relationship between EPS spread and deterministic forecast errors	Spread-skill scatterplot Fitted relationship	EPS spread	Developer + User
EM5: Summarize performance of the probability forecast in parameter space	Continuous Ranked Probability Score; mean and distribution	Probabilistic forecast	Developer + User
EM6: Quantify EPS forecast probabilities for parameter within given bounds	Wilson Score; mean and distribution	Probabilistic forecast	User

Table EP: Measures of (dichotomous) prediction uncertainty in probability space

Purpose	Proposed Metric(s)	Feature of EPS verified	Target Audience
EP1: Quantify deterministic ability to predict event x	Contingency Table Combined % scores: Fraction Correct; Success Ratio; False Alarm Ratio; Miss Ratio	Control Ensemble mean	Developer + User
EP2: Compare the number of EPS members predicting event x with rates of success	Extended Contingency Table	Distribution of EPS members	Developer + User
EP3: Summarize probability forecast errors for event x	Brier Score; decomposition into reliability, resolution and uncertainty	Probabilistic forecast	Developer + User
EP4: Describe probability forecast reliability for event x	Reliability diagram	Probabilistic forecast	Developer + User
EP5: Quantify effect of probability threshold for forecasting event x	Relative Operating Characteristic plot Detection Error Trade-off plot	Probabilistic forecast Threshold criteria	User
EP6: Test long term value of using probability forecasts in decision making	Relative Economic Value score	Probabilistic forecast	User

Table EX: Assessment of performance in extreme conditions*

Purpose	Proposed Metric(s)	Feature of EPS verified	Target Audience
EX1: Quantify the number of extreme events that were predicted within bounds of the EPS forecast	Bounding Box	EPS spread	User
EX2: Quantify deterministic errors between extreme events and maximum/mean/control ensemble members	Error distribution Bias Mean Absolute Error	Ensemble maximum Ensemble mean Control member	User
EX3: Quantify the number of EPS members indicating event	Mean and distribution of member numbers Detection Error Trade-off plot	Distribution of members Threshold criteria	User

* here proposed as reference events exceeding above 95th %^{ile}

VIII ANNEX B - USER CONSULTATION

VIII.1 Overview of the consultation process

The MyWave project aims to incorporate user feedback into its final definition of operational metrics and proposal for an MCS verification system (project deliverable D4.4). The approach adopted for obtaining this feedback comprises 3 stages:

Stage 1: Preliminary survey of potential users in order to establish user types and interest in verification information.

Stage2: Detailed survey of verification requirements for users identified as having an interest in verification.

Stage 3: Review of specific metrics and forms for presentation with users identified as having an interest in specific applications of verification data.

The final outcome from this process is expected to be a set of metrics and associated metadata that can be linked to particular user types and have undergone a period of trial and review.

VIII.2 Initial findings

VIII.2.1 Stage 1

At writing the preliminary MyWave survey[1] has been provided to 68 potential service users to assess their initial reaction to the project and the concept of a wave component of a Marine Core Service. Responses have been received from 35 users. Questions were included that aimed to identify users based on a hypothetical user categorisation presented in MyWave-WP4(UC). From the responses to these questions an ‘in practise’ breakdown of users comprises:

[1] http://www.surveygizmo.com/s3/1299480/MyWave-Preliminary-Survey

- *All Scales Developer-Forecasters*: 7 respondents said they worked with wave information from data generation at both global/large regional scales and coastal scales through to provision of forecasts, and that their data and products were used both for planning and operational purposes. These users were split 70%-30% between commercial and government institutions.

- *Coastal Developer-Forecasters*: 9 respondents said they worked with wave information from data generation at coastal scales through to provision of forecasts, and that their data and products were used both for planning and operational purposes. These users were split approximately 60%-40% between commercial and government institutions.

- *Forecasters:* 11 respondents said they worked specifically on providing forecasts and decision aids and, across the group, undertook an even split of tasks focused on marine operations, hazard forecasting and long term planning (using past climatology). These users were split approximately 50%-50% between commercial and government institutions, with one member of the general public also falling into this category.

- *Decision Makers:* 4 respondents said they generally acted as decision makers and, across the group, undertook an even split of tasks focused on marine operations, hazard forecasting and long term planning (using past climatology). These users were split approximately 50%-50% between commercial and government institutions.

- *Developer-Planners*: 4 respondents were involved in niche model development activities at various scales for planning purposes. These users were split 75%-25% between academic and government institutions.

Of these users 25 expressed an interest in further contact on the subject of MCS verification and were split as 6 All Scales Developer-Forecasters, 7 Coastal Developer-Forecasters, 7 Forecasters, 2 Decision Makers and 3 Developer-Planners.

VIII.2.2 Stage 2

A survey containing more detailed questions regarding user requirements for wave verification[2] was issued on 9th September 2013. Key findings from initial responses (14 users, split as 6 All Scales Developer-Forecasters, 3 Coastal Developer-Forecasters, 3 Forecasters, 1 Decision Maker and 1 Developer-Planner) are that:

- The main requirements for verification data relate to review and intercomparison tasks rather than use in downstream intervention strategies.
- A majority of users would be interested in near-real time monitoring data and downloadable match-up information in addition to review statistics.
- Interactive webpages were considered the best method to deliver verification data.
- Overall wave height, period and direction were considered the most important parameters to verify by all users. A 50-50 split in user requirement was found for verification of more detailed parameters.
- Users considered verification of accompanying wind data as a high priority. Verification for high energy events and a separation of the verification according to wind-sea and swell dominated conditions were identified as important specific aspects of model performance to be tested.
- Quantitative measures of parameter errors were considered to be generally more important than measures of performance for predicting given events, with the exception of high energy storms.
- Where ensemble prediction system verification is conducted, users were keen to see performance cross-referenced against a deterministic forecast.
- Users expressed a preference to see verification statistics referenced against raw observations (i.e. without accounting for observation errors), a distinction made between in-situ and satellite data verification and an effort made to account for sampling and temporal variations within the verification's presentation.

[2] http://www.surveygizmo.com/s3/1306387/MyWave-Verification-Survey

- Metadata describing metrics, observed data used as a reference and quality control procedures should accompany the verification.