MyWave

Proposal for wave verification within a Marine Core Service

Reference: MyWave-D4.4

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	
MyWave version scope : All	
Approval Date : 07 Oct 2014	**Approver:** Andy Saulter
Dissemination level: Project	

DOCUMENT

VERIFICATION AND DISTRIBUTION LIST

	Name	Work Package	Date
Checked By:	Andy Saulter	WP4	07 Oct 2014
Distribution			
	Ø. Saetra (Project coordinator) A. Saulter (WP4) J.-R. Bidlot (WP4) M. Gomez-Lahoz (WP4) Jan Maksymczuk		

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By
0.1	26 Sep 14	all	First draft of document	Andy Saulter	Andy Saulter
1.0	07 Oct 14	all	Document finalization	Andy Saulter	Jean Bidlot, Marta Gomez-Lahoz, Jan Maksymczuk

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABREVIATIONS

ASAR	Advanced Synthetic Aperture Radar
CF	Climate and Forecast
ECMWF	European Centre for Medium range Weather Forecasts
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FTE	Full Time Employee
IOC	International Oceanographic Commission
JCOMM	Joint Commission on Marine Meteorology
MAE	Mean Absolute Error
MCS	Marine Core Service
MERSEA	European FP6 ocean project – precursor to MyOcean
netCDF	Network Common Data Form
OSI-SAF	Ocean and Sea Ice Satellite Application Facility
PC	Production Centre
Q-Q	Quantile-quantile (plot)
QuID	Quality Information Document
Reference data	Data used to verify a prediction
RMS	Root Mean Squared (value of parameter)
(R)MSE	(Root) Mean Squared Error
SI	Scatter Index
TAC	Thematic Assembly Centre
WMO	World Meteorological Organisation

APPLICABLE AND REFERENCE DOCUMENTS

Applicable Documents

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

Reference Documents

	Ref	Title	Date / Issue
DR 1	MyWave-D4.1	MyWave: Definition of experiement plan and resources for MyWave Task 4.1: Identify 'compatible metrics' using remote sensed and in-situ wave measurement baselines	September 2012 / v1.0
DR 2	MyWave-D4.2a	MyWave: Proposal of metrics for user focused verification of deterministic wave prediction systems	October 2013 / v1.0
DR 3	MyWave-D4.2b	Proposal of metrics for developer and user focused verification of wave ensemble prediction systems	October 2013 / v1.0
DR 4	MyWave-D4.3	Estimation of regional observation errors and application to MyWave metrics	December 2013 / v1.0
DR 5	MY02-PQ-CVGWP	MyOcean document MYO2-PQ-CVGWP "MyOcean2 Product Quality Cal/Val Guidelines"	November 2013 / v2.0

I INTRODUCTION

MyWave WP4 seeks to define operational verification methods that can be robustly applied to a wave forecast products provided as part of a Marine Core Service (MCS). The work package has included project tasks that:

- Have assessed observation uncertainty in a regional context and proposed an approach to the application of both this information and sampling uncertainty within verification.
- Have examined the potential to issue consistent verification using a mix of both satellite remote sensed and in-situ observations of the true sea-state as a reference.
- Engaged users of wave model data, in order to understand their needs for verification data and to receive feedback on the proposed verification methods and presentation of the results.

This report proposes a system to deliver verification to accompany operational wave forecasts within the context of a Marine Core Service (MCS). The proposal is based on user feedback on techniques and general requirements for verification, plus a desk study of existing verification processes (e.g. for the MyOcean service). The proposal will discuss the following aspects of the system:

1. Responsibilities of stakeholders involved in the production of verification data, both operationally and for model commissioning.
2. Acquisition of observations and generation of background (model-observation match-up) data for the verification.
3. Options for data processing within the verification scheme.
4. Key metrics and presentation of results.
5. Update cycle for verification data.
6. Requirements for discovery and data access.
7. Review and update cycle for verification scheme.

The intention is that the proposal will form the basis from which a technical implementation of the verification system can be made in any operational follow-on project.

II BACKGROUND TO THE PROPOSAL

This section presents background information that has been critical to forming the proposal.

II.1 User requirements

The process of gathering user requirements followed two stages. In the first, surveys of users were carried out to obtain a high level view of their expectations and requirements for verification data. In the second stage, key representatives of different user communities were presented with a number of verification metrics, presented based on techniques discussed in MyWave-D4.3, and asked for feedback. A summary of the results from this process are presented.

II.1.1 General user requirements

The headline results from user surveys (previously presented in MyWave-D4.2a) were:

- In a Phase 1 survey returned by users, 77% cited verification as crucial information to accompany a forecast service and 40% (14 users) returned a further survey with more detailed questions specifically about verification.
- The main requirements for verification data relate to review and intercomparison tasks rather than use in downstream intervention strategies.
- Interactive webpages were considered the best method to deliver verification data.
- In addition to published verification statistics, a majority of users would be interested in near real-time monitoring data and downloadable match-up information.
- Overall (significant) 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 (such as spectral components).
- 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.

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

II.1.2 Detailed user requirements

A document (reproduced in the Appendix of this report), illustrating a number of metrics proposed in MyWave-D4.2a and using the methods in MyWave-D4.3 for analysis and presentation, was provided to a subset of users representing the following sectors (user types following the schema in MyWave-D4.2a are given in brackets):

- Offshore oil and gas metocean expert (Decision Maker)

- Commercial marine forecast provider (Forecaster)

- Public service marine operations forecaster (All-Scales Developer-Forecaster)

- Public service coastal flood warning forecaster (Forecaster)

- Commercial metocean consultant (Coastal Developer Forecaster)

The feedback process took the form of email and telephone exchanges, from which a number of generic conclusions have been drawn:

1. The users expressed a preference toward simple metrics which gave a clear message without further interpretation. For example one user described Tests C1

and M2 (which provide multiple moment information for particular quantities, see Appendix) as "*seems quite complicated and I am not sure how I would use the statistics*"; another said "*in order to get a feedback from the users, maybe some of them are going to be too complicated to understand and perhaps not so useful for them, as for instance C1 or M2*". On the other hand more than one user expressed a preference for using Bias and Mean Absolute Error (MAE), when communicating their reasons for using a certain forecast to other users, due the simplicity of the metrics. Similarly, Test P1 (which is a simple metric expressing likelihood of a forecast to fall within a given tolerance) was well received, *"looks like it might be a useful statistic to quote to our end users"; "marvellous, answers some of the queries we get"*.

2. Stratifying the data in order to link wave conditions to forecast performance was valued, particularly when evaluating the errors associated with storm conditions. The users liked both quantile-quantile (QQ plots, e.g. Test C2) and Error versus Forecast Range plots (Test R1), "*the first one is a useful plot because looking at overall bias etc. doesn't give the detail of how things vary with Hs level. I tend to use qq plots to give a qualitative feel of bias and standard deviation with varying sub-ranges but your plots look like a more quantitative way of doing it*"; "*really nice plots by quantile*".

3. The use of 'box-and-whiskers' type displays to illustrate confidence levels for the verification statistics (associated with sample size) was neither welcomed or problematic for the users interviewed. One user said "*the mean statistical measure is the one I use. I do not make a good deal of use of the whiskers plot*". Use of idealised performance values provoked more comments, for example "*In some plots it is good to see what the ideal and no-skill results would be. Although in my world, I tend to assume measured is correct and forecast is compared to that. On that basis, the ideal case is getting zero MAE for example, so is obvious*", whilst another user commented "*the [idea that] 100% prediction skill not possible is interesting*". From these comments it would seem reasonable to believe that publishing idealised performance scores would require further explanation to a majority of MCS users.

4. Publishing mapped views of the verification, in addition to area based statistics, was believed to be useful. Comments included "*map based: very useful structure - e.g. could be useful for us to define extent of local models etc.*"; "*definitely good to see stats geographically*".

5. Some further requirements for verification were also suggested in the feedback:

- Parameters verified: "*would like a different measure [on maps], e.g. 95th/99th percentile or maximum of Hs, and perhaps something on period/steepness*".

- Extra statistics: "*one plot that I like to use shows the error distribution. It tends to look a bit like a normal distribution but gives a good feel for the 95% range and any bias characteristics*". It is noted that some metrics had been presented using this type of approach (M3 and R1b), so a clearer presentation or redefinition of these metrics should be considered.

- Model-model verification: "*forecast consistency is also an area of interest, so you can see how much each successive forecast has changed from the previous one. I produce similar stats for this as for overall accuracy against measured data*", "*I also like to show time series plots of measured data and have this overlaid with successive forecasts. This gives a good qualitative feel of forecast consistency too*".

- Long-term archiving of verification: "*tracking of overall MAE, ME is useful and also focussing on the higher sub-ranges to give an indication of improvements through time*".

- Downloadable match-up data: "*I also like to save the actual forecasts (as csv files, say) and measured data to allow post-analysis of specific forecasting issues on an ad hoc basis*".

II.1.3 Recommendations from user requirements

Following the consultations, an MCS verification system targeted at marine users (rather than upstream model developers) is recommended to address the following requirements:

- Verification data should focus on a direct comparison between prediction and reference observation (rather than a corrected result that attempts to account for observation errors). The verification should retain a separation of metrics measured against different observed references (e.g. in-situ and satellite data). Metrics that illustrate forecast consistency (i.e. the amount of change a user might expect from one forecast issue to the next) could also be considered.

- Presented metrics should quantify the verification in real terms (e.g. quantified error, probability of forecast success/failure) rather than as an abstracted skill score. On a

similar note, a requirement for simplicity in the metrics provided should take priority over use of additional data, for example to show verification score confidence levels or 'idealised' model performance.

- Web based publication of verification is the most convenient form for users. Published verification should concentrate on simple metrics requiring minimal explanation. This approach can be complemented if match-up data is available for download by users with an interest in carrying out their own verification.

- Verification data should be archived so that long term performance changes can also be identified.

- The primary requirement for wave verification should be to provide statistics on prediction of (overall spectral) significant wave height and, where possible, wave period and direction. Accompanying wind speed and direction statistics should also be provided where available. In addition to general statistics, verification that stratifies performance based on conditions, or which focuses on the performance of forecasts in high energy storms should be considered.

- Where relevant the verification should include both mapped and aggregated views of some metrics.

II.2 Availability and quality of verification data

The availability of wave observations as a verifying 'truth' has been discussed in MyWave-D4.2a Appendix B. The findings are reproduced as part of the text below.

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. Two observed sources of reference data can presently be focused on for operational verification. '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 (e.g. altimeter, Advanced Synthetic Aperture Radar) 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.

Overall sea-state characteristic parameters, focused on by users and commonly observed by various instruments, are listed in Table 1. Instrument numbers are based on the global network, and in general the number of observations available in specific European sub-regions can be estimated to be (at least) an order of magnitude less than these values. What becomes immediately apparent is that significant wave height is observed in significantly higher volumes than other wave data (note that wind speed data are sampled at a similar volume to wave height).

Table 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
Maximum wave height	In-situ (approx. 20 instruments globally)	

However, even for significant wave height, samples may be limited. For example, using a simple estimate that a verified area will contain 10 observation locations and that data are available hourly with a 95% return rate, the available sample for verification per month is approximately 7000 values. This is a reasonable sample size but, when it is also considered that wave parameters are well correlated on temporal scales of six hours plus, such a sample is more likely to realistically represent a sample of 1000 'independent' events (independence is an inherent assumption in the verification). Where fewer instruments (or

satellite passes) are available the sample could well be reduced to a size that will not produce statistically robust verification.

It is expected that verification offered by an MCS should be consistent in methodology across regional forecast systems. With this in mind, it is recommended that the sampling method aims to obtain a sensible verification sample for the 'lowest common denominator' region in terms of observations availability. Based on knowledge of the in-situ networks in European seas, using a minimum network size of 2 observing platforms should be considered as the lowest common denominator, and would suggest that issuing verification based on (minimum) 3 month data samples is ideal. A similar sample period is likely to be necessary to obtain a sufficient volume of satellite pass data within European regional seas. Although adding a level of complexity to the final verification, the confidence levels associated with these data samples could be estimated using the resampling procedures illustrated in MyWave-D4.3.

A degree of quality control will be required for both in-situ and remote sensed datasets. Bidlot et al. (2002) describe existing quality control procedures for in-situ data used by the Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology intercomparison of operational ocean wave forecasting systems (known elsewhere in this document as the 'JCOMM Wave Intercomparison'). Checks are based on an analysis of observed data time-series and comprise removal of values outside an acceptable physical range, removal of data from faulty instruments (for example by removing all constant records 1 day long or more or based on a manually maintained 'blacklist'), and removal of outliers by comparing individual data values to the deviation from the mean of each monthly data record and from the deviation from one hourly value to the next. For remote sensed data, further procedures must be added to remove observations that might be corrupted by presence of land within the observation swath and where satellite data quality flags indicate that values are questionable. For wind data, Bidlot et al. (2002) note that measurements are not always made at the 10m above sea level standard used by wave models and propose a neutral stability correction to this height in the JCOMM wave intercomparison scheme. Wind speed correction to 10m is recommended for any MCS scheme. One observation quality issue that is more difficult to mitigate is the existence of truncation errors in the data returned from different platforms. The level of truncation should be acknowledged in metadata accompanying the verification.

A further consideration regarding the use of observations within wave verification is the 'representation scale' of observations relative to the forecast models. For the JCOMM Wave Intercomparison, which assesses global scale models for the developer community, a deliberate choice has been made to aggregate observations such that the scale represented by each 'super-observation' in the scheme is equivalent to the process representation scale in a global atmospheric/wave model (order 100km). In previous MyWave reports we have argued for, and assessed data, using an approach in which in-situ measurement scales (equivalent to approximately 20km) should be used as the 'standard', since users will be most interested in verification that shows how the forecast model compares with the in-situ 'truth'. This argument is consistent with feedback from at least one of the interviewed users, "*in my world, I tend to assume measured is correct and forecast is compared to that*". In practise this would mean using the in-situ observations in their raw form, super-observing remote sensed data toward the 20km scale, and using the most appropriate scaling for model data (for example, the representation scale of a 6-8km atmosphere/wave model is approximately 20-30km).

Following this discussion, it is recommended that:

- Verification samples are based on (at least) 3 month match-up samples of model and observation.
- Quality control procedures are defined based on the template established by the existing JCOMM wave intercomparison scheme, and updated to make use of quality flags provided with remote sensed data.
- In-situ platforms measurements are used as a standard for the 'representation scale' adopted within verification.
- Metadata provided with the verification should include: observation source(s) and instrument type(s); number of data returned and rejected (by platform/instrument); link to quality control procedure and platform/instrument 'blacklist' documents; information on data super-observation methods as appropriate.
- Subject to a review of 'in practise' data volumes, the use of resampling techniques to estimate sample based confidence limits for the verification should be considered.

II.3 Methods for MyOcean verification

The discussion in this subsection is based on conversations with staff working on the MyOcean verification work package, MyOcean document MYO2-PQ-CVGWP "MyOcean2 Product Quality Cal/Val Guidelines" and a review of the MyOcean published validation pages at http://data.ncof.co.uk/calval/index.html. The methods adopted in MyOcean are an important consideration for this proposal, since it is assumed that the most likely route by which wave data will be published as part of an MCS in future will be through integration with the MyOcean service under the Copernicus programme.

As background, verification development within MERSEA and then MyOcean has been targeted at the upstream model developers and based on establishing common ground for performance measurement. MyOcean2 verification is regularly published: for operational data the data are available via verification webpages; verification from model trials is published within a Quality Information Document (QuID) that can be downloaded from the product catalogue page. The procedures developed are agreed and modified by MyOcean Production Centres (PCs) and a central verification team. The latter party are responsible for the publication of operational data on the MyOcean webpages and to ensure that data are updated on schedule and presented in a consistent manner.

Pertinent details of the MyOcean operational verification procedure are as follows:

- Webpage reports, comprising verification plots and accompanying text are updated quarterly.
- The data sample used in each report update comprises 12 months of data.
- Plots provided are an overall summary of performance from the last 12 months for the PCs overall region and relevant sub-areas, and a time-series of daily verification statistics for each region and sub-area (Figure 1).
- Core verification is based on a limited set of agreed common parameters and metrics (Bias and Root Mean Square Error as primary statistics, with normalised standard deviation and correlation as optional secondary statistics for construction of Taylor diagrams). However, individual PCs can request publication of further metrics/parameters if particularly relevant to their local user base.

- The central verification team is responsible for filtering the verification data provided in order to maintain a clear, simple view of the verification for users. For example, PCs may verify at multiple lead times, but the MyOcean website will only visualize verification at three lead times in order not to clutter plots.

- Presently the MyOcean visuals do not include a long term view of the annual performance verification, e.g. changes in bias/RMSE over a several year period.

- PCs are responsible for providing the verification team with a quarterly update of the daily verification statistics as a standardized netCDF file (format details are described in MyOcean report MYO2-PQ-CVGWP) and any update to the report text (e.g. major updates to observations used or quality control procedures). Report text is not particularly verbose (for example no commentaries on results are provided) in order to minimise update requirements.

- PCs are responsible for their own observation acquisition and quality control procedures (e.g. use of MyOcean Thematic Assembly Centre observations is not enforced) and any communications with the TAC regarding consistently poor observations.

In an envisaged system, where oceanographic and wave products are all released within the same Copernicus MCS, wave verification and ocean verification should be harmonised in order to maintain the consistency and simplicity of presentation criteria established within MyOcean. Nevertheless, dealing with differences in the nature of the verifying observations for each system is likely to require some expansion in scope of the verification procedure. For example, samples of wave data from individual days will not be sufficient to generate daily mean statistics as in the MyOcean case. Since a number of wave PCs are not reliant on data assimilation to generate model starting conditions, the existence of infrastructure to generate match-up samples and provide observations quality control cannot be taken for granted in the wave verification process. Therefore the benefits of using a central observation provider or verification data analysis team might be usefully explored.

In addition, the integration of wave verification to the existing programme MyOcean would provide an opportunity to share methods in order to drive improvements, particularly in view of the user focus that has formed the background for verification development within MyWave. For example, a number of the simple wave metrics endorsed by users (QQ plots, ‘Probability within’ metrics) are equally applicable to ocean data as to wind and wave

parameters. The identified user requirement for simplicity in the verification system could also be extended to creating more 'human readable' webpages than are presently available through the MyOcean portal. In order to incorporate these ideas, the proposal for wave verification will be set above the lowest common denominator of the MyOcean and MyWave schemes.

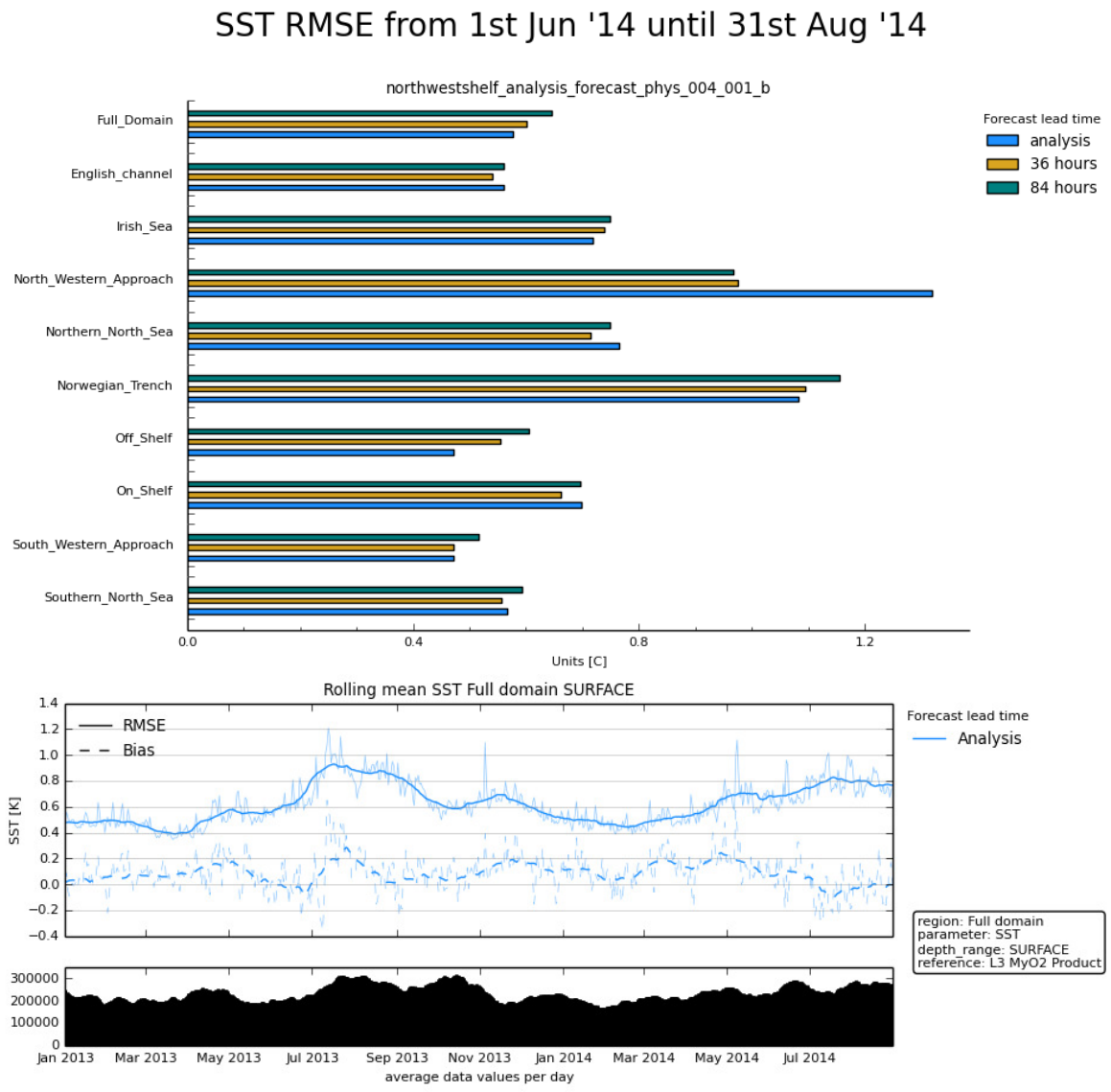


Figure 1. Example core verification plots from MyOcean webpages (link from http://data.ncof.co.uk/calval/index.html). Top panel, annual RMSE for Northwest Shelf sea-surface temperature. Bottom panel, daily mean Bias and RMSE time-series for Iberian sea-surface temperature over full model domain.

III PROPOSAL FOR MCS WAVE VERIFICATION

Based on the evidence presented in Section II, this section provides a proposal for a wave verification system. The proposal uses the key assumption that future delivery of wave data within a MCS will be in coordination with existing MyOcean services, under the Copernicus programme. A further assumption has been made that, initially, the wave products offered and verified under a MCS will be deterministic (i.e. not ensemble data). Although noting areas for future development, the proposal concentrates on procedures necessary to deliver an initial (V0) operational verification system.

III.1 Reporting and update cycle

The process of reporting wave verification should fall in line with the existing MyOcean process. Specifically, verification reporting for trials of new systems should be incorporated into Quality Information Documents (QuIDs) and updated to accompany new model releases. Operational verification should be reported via MCS webpages and updated on a quarterly basis. A proposed addition to the MyOcean methodology is to track the evolution of quarterly summary statistics in order to trace long term changes in operational system quality.

III.2 Workflow and responsibilities

Figure 2 presents a schematic of the workflow necessary to generate and publish operational verification data. The proposed scheme has six steps:

1. Acquisition and preliminary data processing of reference data, e.g. in-situ or remote sensed observations.

2. Generation of ‘Level 1’ verification data - match-up data in which all available reference and forecast data are paired based on time and space collocation criteria.

3. Generation of ‘Level 2’ verification data - match-up data where match-up pairs are flagged for rejection from the verification statistics, according to quality control procedures.

4. Generation of 'Level 3' verification product - 'granular' statistic data, i.e. verification data at (or above) the maximum resolution to be used in verification reporting.

5. Generation of 'Level 4' verification product - 'aggregated' statistic data, i.e. verification data at summary reporting levels (e.g. annual statistics).

6. Publication, i.e. verification plots via MCS webpages and generation of report metadata based on methods in stages 1-5.

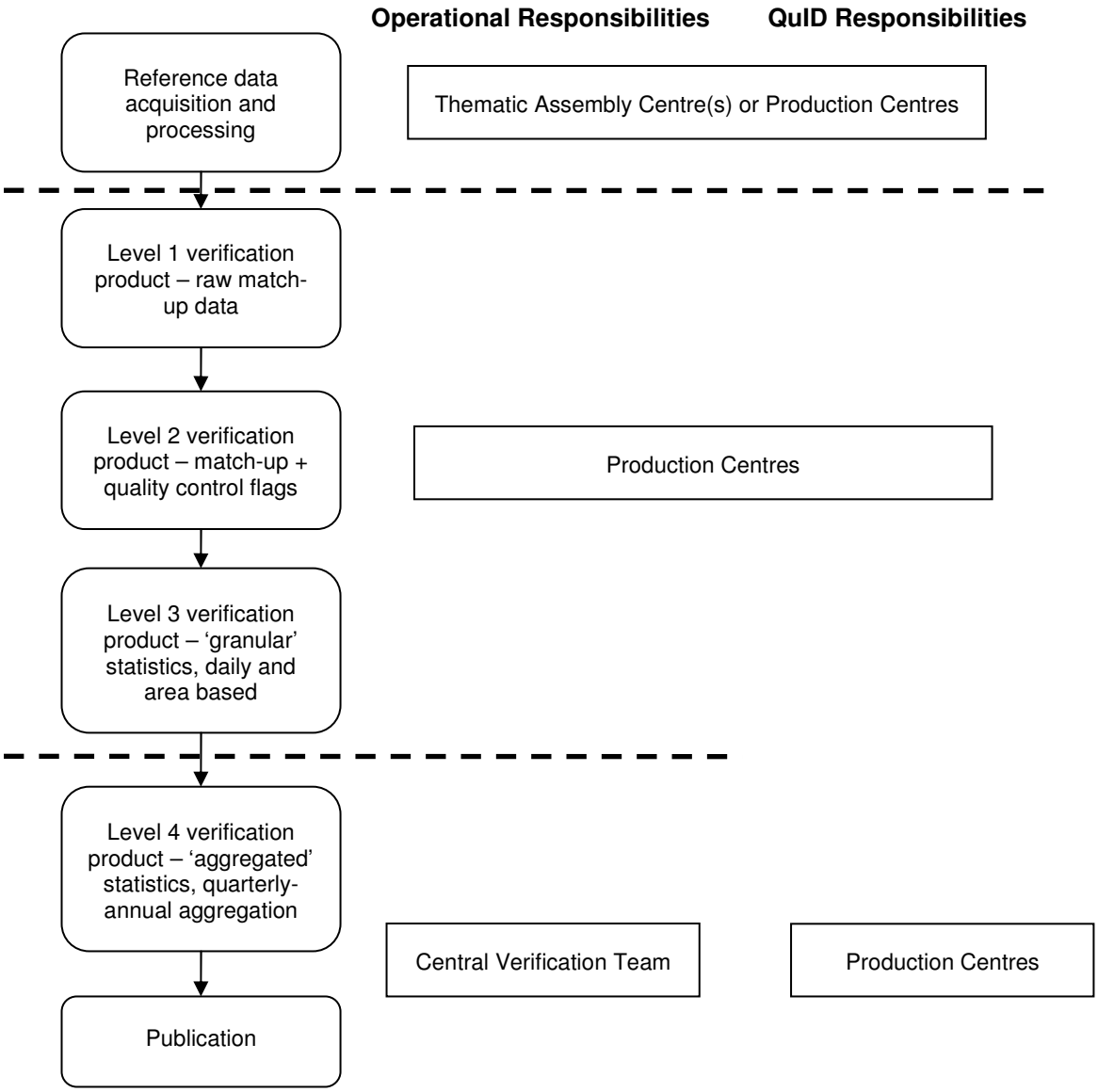


Figure 2. Schematic for proposed wave verification workflow and responsibilities.

In MyOcean, responsibility for the equivalent of stages 1-4 are taken by the Production Centres (PCs) and result in generation of netCDF data files comprising (granular) daily statistics for regions and their sub-regional areas of interest. The responsibility for stages 5 and 6 are taken by a central verification team, with the input to report metadata and text provided by the PCs.

Adopting similar responsibilities for wave verification would be effective and logical, considering that PCs would need to follow through stages 2-6 as part of any verification provided within a model QuID. However, there might be a significant overhead for PCs in developing infrastructure required to gather all the necessary observations. This overhead is envisaged since, as data assimilation is not an essential component of a wave forecast system, some PCs may not have developed operational systems for observation download and match-up similar to their ocean modelling equivalents. However, in an MCS that includes a wave Thematic Assembly Centre (TAC) for observations, it is envisaged that this (these) centre(s) could take on responsibility for gathering the observation data for wave verification and presenting the data in a standard format that could be more easily dealt with by the PCs. This structure would reduce the wider overhead on data acquisition and enable standardisation in the approach to observation quality control.

III.3 Observations acquisition and match-up data processing

Three sources of reference data for verification are identified:

1. In-situ measurements from wave buoys and other platforms
2. Remote sensed observations, including from altimeter (wave and wind), scatterometer (wind) and Advanced Synthetic Aperture Radar (ASAR, for waves, subject to agreement on the parameters to be verified and data processing scheme for the observation).
3. Forecast model data, for consistency analyses (e.g. comparison of the day 5 forecast to forecasts at lead times of 4, 3, 2, 1 days).

It is anticipated that candidate wave PCs will presently have a varied level of infrastructure in place to obtain and process in-situ and satellite observed data. In order for the MCS to mitigate this, one option is that a demonstrated observations gathering and processing

infrastructure is included as a qualifying requirement for PCs wishing to provide data to the waves service. The other is to assist the process by centralising the production of 'verification ready' observation datasets through the service TAC. This second option has the advantage of better traceability (since the observation processing is limited to a very small number of centres) and would also build on some established infrastructure. For example, ECMWF already provide a 'best endeavours' service to key wave modelling centres worldwide by assembling and quality controlling a monthly release of in-situ observations for verification (Bidlot et al., 2002, 2007) and have well established quality control procedures in place for altimeter data. Near real-time and longer term validation procedures are adopted by the EUMETSAT Ocean and Sea Ice Satellite Application Facility (OSI-SAF). The DUE GlobWave project (http://globwave.ifremer.fr/) has established a capability for provision of readily useable satellite observation products, using a consistent and standardised quality control procedure and data format, at IFREMER. Polling members of the MyWave consortium, it was strongly felt that observation quality control procedures (beyond existing near real-time protocols) needed to be put in place. The preferred option was to use a centralised process in order that wave verification observations were of a consistent quality level. Opinions were divided as to whether a TAC was the best route to this, or whether strict adoption of agreed procedures was undertaken at the PCs.

Common to either option is the requirement to define rules for observation data quality control, model-observation match-up method and choice of the representation scale that the verification will centre on (as discussed in Section II.2). These will need to be established through agreement between the PCs (TACs if used) and central verification team. For observation quality control, procedures underpinning the JCOMM Wave Intercomparison and GlobWave 'Pilot Wave Forecast Verification System' can be used as an existing template. The match-up process would be expected to be the responsibility of the PCs and should be consistently defined as either a 'nearest neighbour' or 'interpolated data' method[1]. In addition to agreeing a consistent representation scale, the issue of whether this additional processing is carried out by PCs or TAC(s) needs to be addressed.

A noted user requirement was for data used in verification to be made available for download, so that users could generate their own metrics specific to their particular tasks. It is assumed that this request would correspond to the 'Level 2' verification data produced by

[1] Consistency in this case refers to the method applied by different PCs. In practise the match-up procedure used for different observation types (e.g. satellite, altimeter) may differ for valid scientific reasons.

PCs, i.e. comprise model-observation match-up data and associated quality control flags. Two issues are noted: 1) that this essentially creates an extra product from the service, which would necessitate a data standard and associated documentation (i.e. extra delivery costs) and, 2) that a number of in-situ data sources available for verification are proprietary (e.g. oil and gas platform data from the North Sea) but, whilst not be available for onward publication, would be considered essential to a successful verification scheme. In the latter instance a download verification dataset would not be consistent with the published metrics. In light of these issues it is proposed that the download option is not placed within scope of a V0 verification scheme.

III.4 Aggregation of wave data

The present MyOcean verification scheme makes use of a two stage aggregation process. The first, 'granular', level established by PCs produces verification statistics at a regional and sub-regional level (e.g. Northwest Shelf 'full domain', 'south west approach' sub-region) with a daily temporal aggregation. The quarterly update of these daily time-series is 'aggregated' by the central verification team to generate annual statistics. Both granular and aggregated verification are published.

The number of verification data available from model-observation match-up per day is likely to be too small for daily wave verification to be robust. Indeed, the recommendation in this proposal is that the minimum data sample period for any published statistics is set to 3 months. However, for purposes of consistency with the MyOcean verification production method, provision of statistics at a similar level of granularity (i.e. a statistic value for each day) from wave PCs is feasible. This approach would require that each published 'daily' value is based on a rolling 3 month aggregation of data.

There are two choices to the aggregation method. In the first option, the daily values provided by the PCs are generated direct from a 3 month (rolling) match-up sample. In the second option, the PCs provide daily statistics based only on the daily data sample and the aggregation is dealt with by the central verification team combining the daily statistics. This second option is achievable for the core metrics recommended by this proposal (Section III.5), which can easily be aggregated so long as sample size information is provided alongside the statistic. Option 2 provides the most flexible method. In either case, daily time-series should provide an excellent resource for users, since they will quantify seasonal

variability in the verification (which is a recognised feature of wave model performance in the variable climate of mid-high latitudes). Aggregation of the daily statistics into annual values, consistent with the MyOcean scheme, is also recommended and can be accomplished using either option.

Following the expressed user preference, one area of separation in the data should be between reference types, i.e. separate statistics are available for the comparison between model and in-situ data, model and satellite data, model-model consistency data.

In addition to daily statistics it is also proposed that the PCs maintain and provide a set of rolling annual verification products within the quarterly release, specifically quantile-quantile (QQ) statistics and error data stratified by forecast value. These have been identified as highly useful data by users, but cannot be sensibly reproduced using a daily-to-annual aggregation method.

III.5 Core metrics for deterministic model verification

The following metrics are proposed for a V0 wave verification system, based on user requirement and consistency with MyOcean metrics. Daily data files should include (for each region/sub-region, lead time, parameter and reference type):

- Number of match-up data[2]
- Mean reference value
- Mean squared reference value
- Mean model[3] value
- Mean squared model value
- Mean value of model times reference
- Mean Absolute Error (model-reference)

[2] Within MyOcean verification documentation the number of match-up data are termed ‘data values’.
[3] Within MyOcean verification documentation the term ‘product’ is used to refer to the model data.

- Probability of model-reference error within threshold(s)

Statistics provided in this form enable aggregation and construction of Bias, RMSE, MAE and 'Probability within' metrics; also Taylor diagrams and Scatter Index metrics.

Annual data files should comprise (for each region/sub-region, lead time, parameter and reference type):

- Number of match-up data
- Quantile values (every 1% to 95%, then every 0.1% to 99.9%) of reference
- Quantile values (every 1% to 95%, then every 0.1% to 99.9%) of model
- For each model prediction bin[4] (to generate error through range data):
    - Number of match-up data
    - Model-reference bias
    - Model-reference mean squared value
    - Model-reference MAE

III.6 Parameters for verification

Operational verification at V0 should focus on the most readily available data, which are also easily recognised by users. The parameters identified are significant wave height and period (period type is dependent on local observation protocols, but different periods should be treated as separate parameters), wind speed. Other parameters (e.g. wave and wind direction parameters, wave spectra) should be considered within QuIDs, as availability of observations for product verification allows.

Metrics for combinations of parameters (e.g. wind speed and direction, significant wave height and period, steepness) are considered outside of scope for a V0 system, but should be a consideration for both V0 QuIDs and future verification system development.

[4] In principle, data following this stratification could be provided within the daily statistics, but it would be expected that many days would include a null return for at least one prediction bin – rendering the daily data somewhat meaningless.

III.7 Additional metrics and future requirements

Within the MyWave project the use of a large number of metrics, beyond the core set proposed in Section III.5, has been explored. For V0 verification it is proposed that the use of additional metrics is a choice made by PCs within the process of producing the regional system QuID. It is noted that the MyOcean service provides a user feedback process through which requests for revisions to the verification data, including for new metrics, can be made. It would be expected that this process is put in place for wave services and is used to inform updates to the core metrics (see section III.11).

Model lead time 'consistency' data should be within scope for the V0 service at regional and sub-regional aggregation levels. However, it is noted that these statistics would also lend themselves well to mapped visualization, since a match-up can be made daily at every point within the model domain. It is suggested that the utility of this type of visualization be explored with users for later versions of the verification system. A mapped view of verification by platform could also be considered for in-situ data.

This proposal has assumed that ensemble wave products are initially out of scope for a waves service. Should such products be added to the catalogue, the verification programme would need to be expanded in order to cover the use of ensemble data. Ensemble core metrics should include:

- Deterministic core metrics as for V0, applied to ensemble control and mean product.
- Metrics for spread skill comparisons (e.g. daily mean spread)
- 'Reliability' score for predefined threshold(s) (e.g. Continuous Rank Probability Score, Brier Score)

More details of ensemble metric types are provided in report MyWave-D4.2b.

III.8 Further options for statistical processing within the verification

The MyWave project has explored three methods to contextualise verification results, derived from the direct comparison of model and an observed reference (see report MyWave-D4.3):

1. Re-sampling to help understand the effects of the sample used.

2. Generation of idealised verification data to estimate target performance levels.

3. Generation of naïve prediction verification data to define a low performance boundary.

Method 2 is reliant on the data sample being sufficiently large that randomly drawn data values, representing observation errors, can be used to generate a statistically robust 'idealised observation' dataset. For the production of daily statistics, under aggregation option 2 in Section III.4, this criterion is unlikely to be met.

Similarly, aggregation option 2 would preclude the use of resampling techniques by the PCs. For low samples of data, the results of resampling (particularly the lower/higher percentile values of the resulting metrics) are unlikely to be stable enough to enable a robust aggregation (for the purpose of producing confidence limits for quarterly or annual statistics) later in the verification process. An option might be for the central verification team to resample daily statistics within their aggregation process[5], although this would be dependent on the consistency in the size of daily data samples.

Method 3 was not identified by users as a 'must have' for their verification.

The application of these statistical processing methods is therefore dependent on determining where in the verification process data aggregation (to the 3 month sample size) takes place. It is also worth bearing in mind the extra level of complexity introduced to users in interpreting data processed using these methods. Overall, it would seem sensible that the use of any statistic processing at V0 is initially constrained to voluntary implementation by PCs within QuID verification.

III.9 Verification data format and metadata

Following the existing MyOcean standard, the proposed file format for wave verification statistics is (CF compliant) netCDF. NetCDF (Network Common Data Form) is an open standard, self-describing binary data format that is in common use in the climate science and oceanographic communities. This is also the common format for products across the MyOcean service. Details of the conventions for verification file naming, style, dimensions, coordinate variables, statistics variables and attributes, global attributes and file metadata

are given in MyOcean report MY02-PQ-CVGWP. It is anticipated that wave verification data can be issued in keeping with these protocols.

In addition to data, the MyOcean service also requires production centres to supply summary report text. This is generally static content, such as in the example below (from the Mediterranean PC):

> "*Statistics are computed on differences between analysis and forecast daily means and in situ observations from Argo floats and moored buoys and daily MyOcean SST satellite L4 regional products. Detailed results are also available from http://gnoo.bo.ingv.it/mfs/myocean/evaluation.html and http://gnoo.bo.ingv.it/myocean/calval/. Temperature and salinity observations from moored buoys are independent observations. These buoys are unevenly distributed in space (http://gnoo.bo.ingv.it/myocean/calval/) and have a consistent number of observations only in the first 10m of the water column. Results may be influenced by sparse data cover.*"

For wave data it is proposed that similar text would be provided alongside other metadata underpinning the verification. Ideally, the following items should be incorporated with the report/metadata:

- 'Plain language' explanation of the metrics.
- Details of quality control and other data reprocessing methods (e.g. rescaling of data using 'super-observations').
- List of instruments/platforms/satellite mission(s) forming the observation data sample, and the known precision of the data (e.g. to identify truncation errors).
- List of 'blacklisted' platforms/satellite missions.
- Number of observations in 'Level 2' verification match-up sample and number of rejections.
- (for in-situ data) Number of observations and location by platform.

[5] This approach would be equivalent to applying a block bootstrap, with the block size set temporally at 1 day.

III.10 Discovery and data access

One purpose of the MyWave user feedback process was to assess whether users work with verification data in sufficiently different ways to necessitate a level of data discovery to be introduced to the published verification. In general however, users were found to focus on the same group of simple metrics, corresponding to the core metrics described in section III.5. For a V0 verification system it would therefore seem unnecessary to add any architecture for verification data discovery beyond the web link process used by MyOcean. Indeed, if the MCS provides regional wave products consistent with the existing MyOcean catalogue, then wave parameters could simply be added to the present webpage structure.

With regard to web publication, one user consideration not in the existing MyOcean pages is for simpler terminology to be used in the web links. Presently these use PC product identifiers and parameter acronyms, which are likely to be off-putting to non-technical users.

It has not been ascertained whether daily, area aggregated, statistics would meet wave users' expressed requirement to access data underpinning verification (or whether this is restricted to the Level 2 verification data). In this proposal it is expected that such a download facility, if required, would be created as part of the system beyond V0, since a similar should then be expected for ocean data.

III.11 Review cycle and governance structure for the verification scheme

As with the models used to generate the data, the verification scheme should not be considered a static entity, but as a programme of the work to be regularly reviewed and updated. This should be an annual process and consider the following aspects of the system:

- Availability/retirement of observation platforms.
- Availability of new parameter types in observations.
- User feedback and requirements for development of the system (e.g. via an MCS forum).
- PC/central verification team requirements for development of the system.

- Webpage monitoring.

It is proposed that governance follows the existing MyOcean structure, in which the central verification team takes the overall lead for operational verification. Governance comes from the collective group of PCs and central verification team, in response to both internal drivers for change and new user requirements (acquired through the MCS feedback service).

IV IMPLICATIONS FOR PRODUCTION CENTRES

Adopting the proposed method will have implications for PCs. Providing details of individual impacts to modelling systems and requirements for resourcing in individual PCs is outside the scope of this project and would be premature, but the likely implications can be assessed by examining the state of readiness at the Met Office as an example.

The Met Office has significant infrastructure for data acquisition and processing in near real-time and is developing the technology to match-up observations and model during regular daily 'update runs' its wave forecasting system. Other service providers, using data assimilation or nowcasting procedures in their service portfolio, will have similar, or more advanced, capabilities. However, the Met Office waves team makes heavy use of quality controlled observations, provided via the JCOMM Wave Intercomparison scheme in verification, in preference to near real-time data. On this basis, whichever observations acquisition and data processing option is chosen, a (limited) overhead would be incurred in adopting the MCS verification procedure for match-up and observations quality control as part of Met Office 'business as usual'.

Wave verification at the Met Office is presently an internal procedure using local file formats and visualization conventions. Therefore a significant piece of work would be necessary to generate daily and annual verification statistics and quarterly metadata adhering to the proposed data formats and standards.

Similarly, the team has no major driver to produce regular documents equivalent to MyOcean QuIDs. Establishing this process as part of business as usual would also be considered as an overhead on participation within an MCS.

As an indicative estimate (based on provision of a single regional product), the Met Office would expect to require up to 1 full-time employee (FTE) resource to create the V0 verification data generation procedures and provide an initial QuID. Resource of approximately 0.3 FTE would be likely required to support the QuID and operational verification process in subsequent years.

V SUMMARY

This report has documented the drivers for and proposal of a wave verification scheme to accompany delivery of (deterministic) model wave forecast products within a Marine Core Service (MCS). The focus of the proposal has been on the necessary structures and procedures for delivery of an initial (V0) operational verification scheme.

The proposal has been based on a number of assumptions, of which the most critical is that that future delivery of wave data within a MCS will be in coordination with existing MyOcean services, under the Copernicus programme. This assumption provided a major driver for the form of the proposal since, for purpose of consistency, the most efficient way to merge ocean and wave verification would be for the wave verification programme to adopt a number of the existing MyOcean procedures: including production method, core metrics, data standards and publication methods.

In the proposal, the division of responsibilities for verification data production has been made along the same lines as for the MyOcean process. Production centres (PCs) will be responsible for model-reference match-up and generation (and quarterly release) of daily statistic files, with a central verification team (potentially the existing MyOcean team) responsible for further data aggregation and publishing of operational statistics. PCs will also be responsible for verification within product Quality Information Documents (QuIDs). One option that departs significantly from the MyOcean role assignment would be to place responsibility for observation data acquisition and quality control in the hands of a Thematic Assembly Centre(s), in order to provide a centralised resource for these data.

In terms of operational metrics, user feedback (including a requirement for simplicity) has led us to propose only a limited extension to the data presently published by MyOcean. Most notable extension is provision of (annually aggregated) quantile-quantile data and error estimates stratified by the conditions being predicted. Additionally, long term tracking of core metrics is proposed, in order to provide users with a long terms view of progress in improving product accuracy. Other metrics and data processing methods, studied during the course of the MyWave project, are expected to be best suited to verification carried out within generation of QuIDs.

It is anticipated most wave PCs have a level of infrastructure in place which can be adapted to meet the requirements of this proposal. It is noted that some (best endeavours) architecture for global wave model verification already exists and is used by a number of European centres. However, it is expected that, in order to meet the level of detail and standardisation that would be required to integrate wave verification into the existing infrstrucutre used by MyOcean, additional resourcing will be required (both for set-up and ongoing support) at the majority of candidate PCs.

At this stage, the proposal should be considered as an initial position on wave verification within an MCS. Other options to deliver such a system exist and it is envisaged that details of how the scheme might be run would need further review and modification at the technical implementation stage. These details would be best reviewed and agreed once a governance structure for the scheme is established. In particular, it would be expected that this group would provide the best forum to agree details on methods to acquire, quality control and process observations for verification.

VI REFERENCES

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

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.

VII APPENDIX – VERIFICATION METRICS DOCUMENT PROVIDED TO USERS FOR FEEDBACK

Purpose of document

This document is intended as an aid to discussions regarding the utility of verification statistics identified in MyWave WP4. The aim is to show users examples of various statistics and receive feedback on the following:

- Would you use this statistic to help your use of model data?
- Is the statistic and its presentation understandable?
    - Are the contextual information useful?
    - Is the statistical uncertainty information (plumes, box-and-whiskers) useful?
- Are there alternative methods of visualization that you would like to see?
- Are there particular metrics that you would like maintained over the long term?

Area based statistics

The following examples are proposed statistics to describe deterministic model performance in given regional sea areas. As an area average these data can be generated from either an amalgamation of in-situ platform data or satellite measurements. Data to contextualise the verification is based on an assumed knowledge of scale and model for observation errors and a prescribed 'no skill' forecast scenario.

General points on the statistics

In the majority of cases the statistics are derived from a 'bootstrap' ensemble of matched data samples, so that the statistic value will have a range of outcomes. These results are plotted using 'box and whiskers', plume or 'cross-hairs' symbols.

In a number of cases the model-observation verification is contextualised by an 'idealised verification', derived by assuming that the model represents the true state and the statistic is entirely down to observation errors, and a 'no-skill prediction', derived by using a random sample from the model data as a prediction. The idealised data are represented by a green plume/symbols and the no-skill data by an orange plume/symbols. The model-observation data are represented by blue symbols.

For each example a short description of what the plot should tell the user has been given. The aim of this is to help the discussion process.

Details of symbology

'Cross-hairs' show the mean statistic value and 5-95% range that might be expected as a result of sampling variability.

A 'box-and-whiskers' display is used to indicate the likely variability of the model-observation statistic associated with sample size. For the box-and-whiskers the centre circle defines the mean statistic value, the inner box lines define the inter-quartile range, the outer box lines define the 5-95% range and the whiskers define the 1-99% range for the statistic.

A green 'plume' shows an estimate of 'ideal performance', i.e. the value the statistic would take if the model forecast was perfect and the only contribution to the statistic came from observation errors. The plume shading defines inter-quartile range, 5-95% range and 1-99% range for the statistic.

An orange 'plume' shows an estimate of 'no-skill performance', i.e. the value the statistic would take if the model forecast was a random estimate based on the model climate. The plume shading defines inter-quartile range, 5-95% range and 1-99% range for the statistic.

Source data for examples

The examples in this document have been constructed from a dataset comprising 3 months of 6-hourly match-ups of in-situ observations and model data for sites in the northern and

central North Sea. These data were sourced from the JCOMM wave forecast intercomparsion dataset maintained by ECMWF.

A block bootstrap was applied to the data in order to evaluate sample effects (i.e. the plumes and box-and-whiskers in the plots). Blocks were defined as 24 hour periods for 3 150km^2 areas containing platform data, with a random draw process used to select data such that an equal sample was provided by each block. The bootstrap ensemble comprised 1000 members.

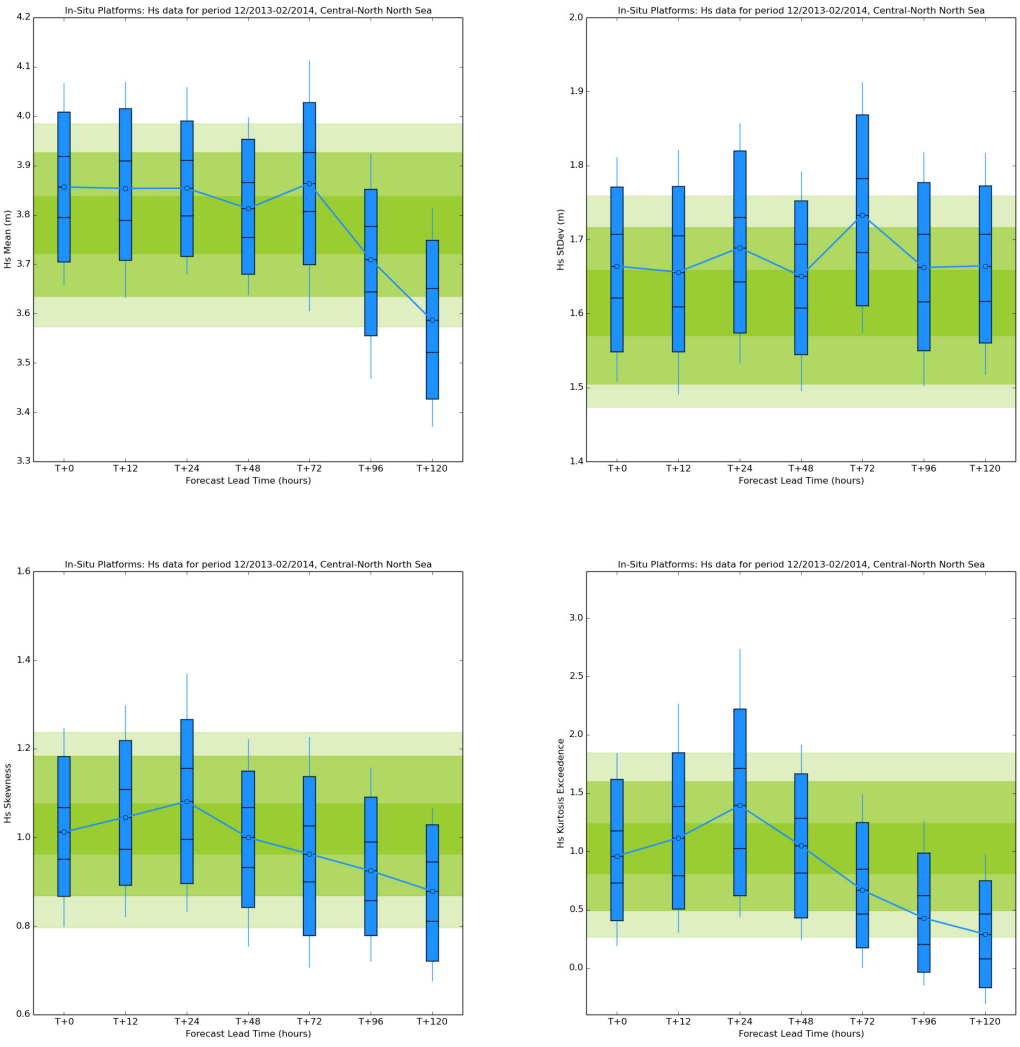
Test C1: Does my model reproduce general features of the reference climate?

Verification: Comparisons of climate statistic moments

What does this tell me?

The presented '4-up' view shows forecast lead time changes in the modelled statistics representing 1st, 2nd, 3rd and 4th moments of significant wave height distribution. The green plume shows the observed moments. Good performance is identified when the plume and box-and-whiskers overlap for all 4 plots.

In this example, performance at short lead times is good, although the model slightly overestimates conditions. Conditions are underestimated and the shape of the distribution is less well specified at longer lead times.



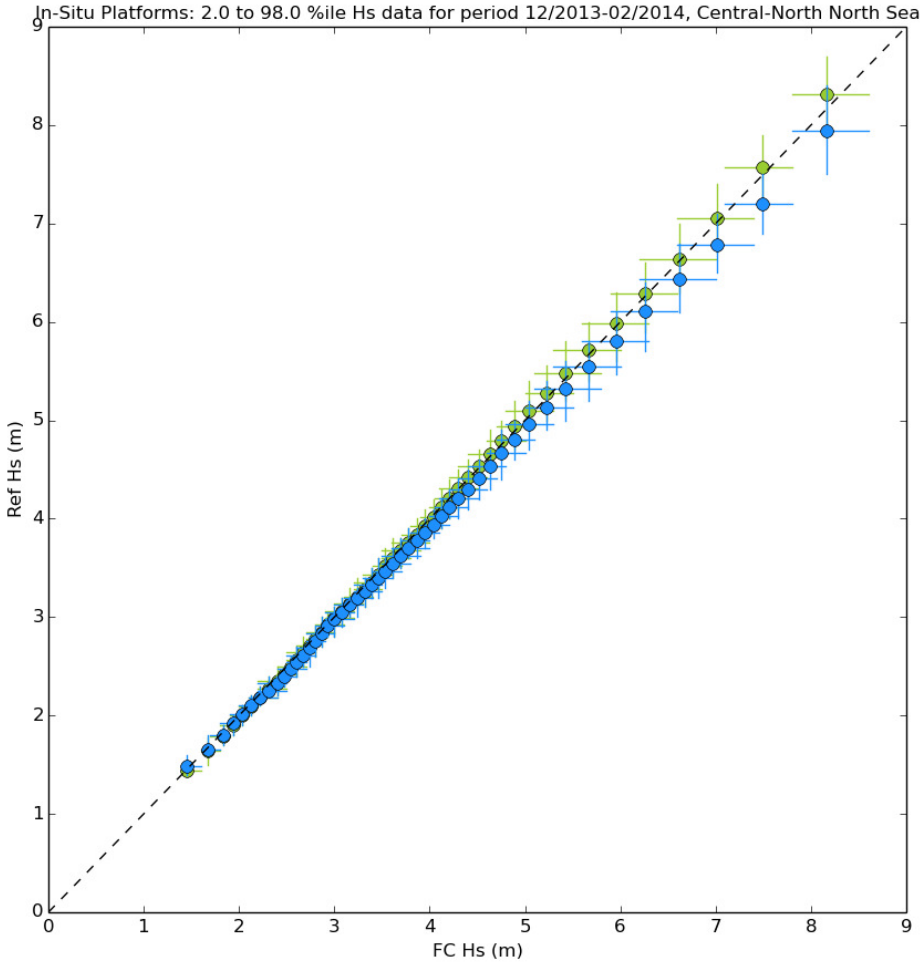
Test C2: Does my model reproduce details of the reference climate?

Verification: Quantile-quantile (Q-Q) plot

What does this tell me?

The plot compares frequency distributions of observed and modelled occurrence of the given variable. Data points deviate from the 1:1 line when, for a given percentile of the frequency distribution, a higher value occurs in either the observed or modelled data. For example a constant offset of points from the 1:1 line will indicate a systematic bias, whilst a curve away from the 1:1 line indicates that the distribution is under- or over-sampled.

In this example the model-observation points (blue) show excellent representation of the climate up to 5m Hs (80th percentile) and a slight overprediction at higher percentiles. The difference in location between blue symbols and green symbols (representing the ideal statistic) suggest that the overprediction becomes significant above the 90th percentile.



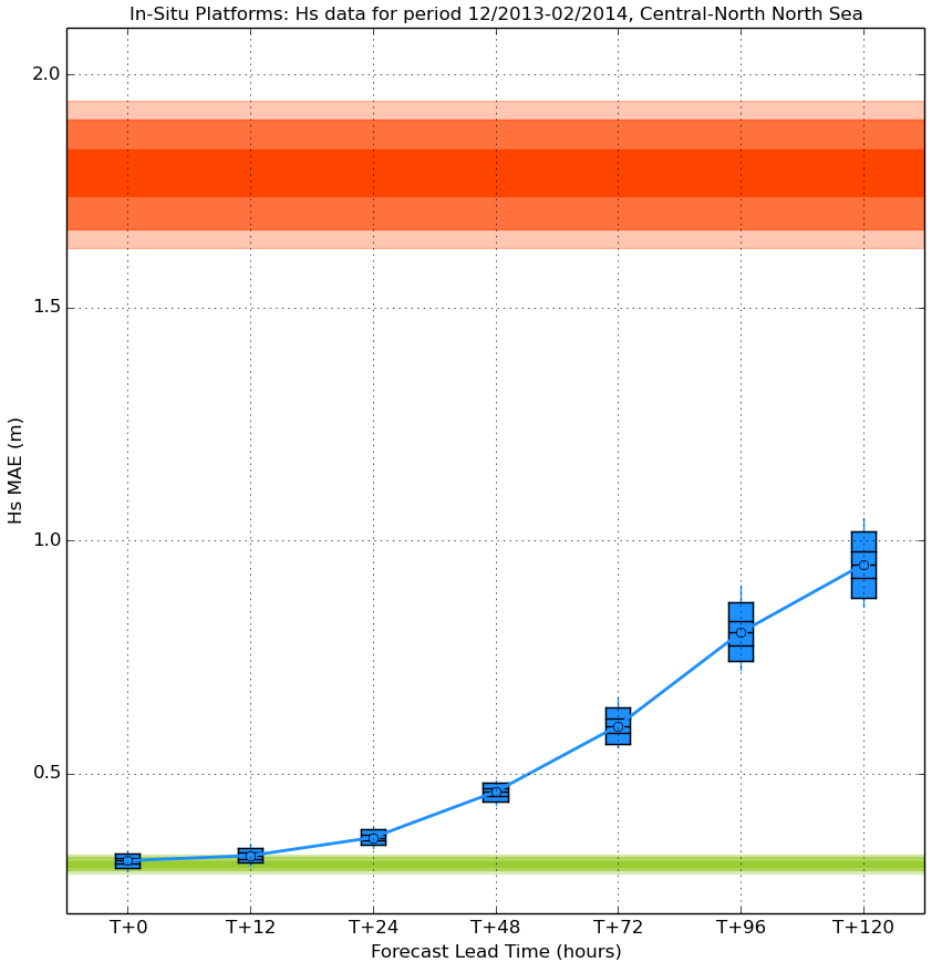
Test M1: Quantify the scale of prediction errors

Verification: Mean Absolute Error (MAE) plot

What does this tell me?

The plot shows the average (absolute) error between observed and modelled instances of the given variable over the data sample. MAE will increase with (absolute) bias and as the model's ability to replicate the temporal signal of the observations decreases.

In the example shown, MAE at short lead times falls very close to the idealised case, i.e. the majority contribution to model-observation errors is expected to be a result of observing error. At longer lead times the model errors are a number of times higher than the observed error. However, even at 5 days ahead (T+120) the model is significantly more skilful than a 'random guess'.

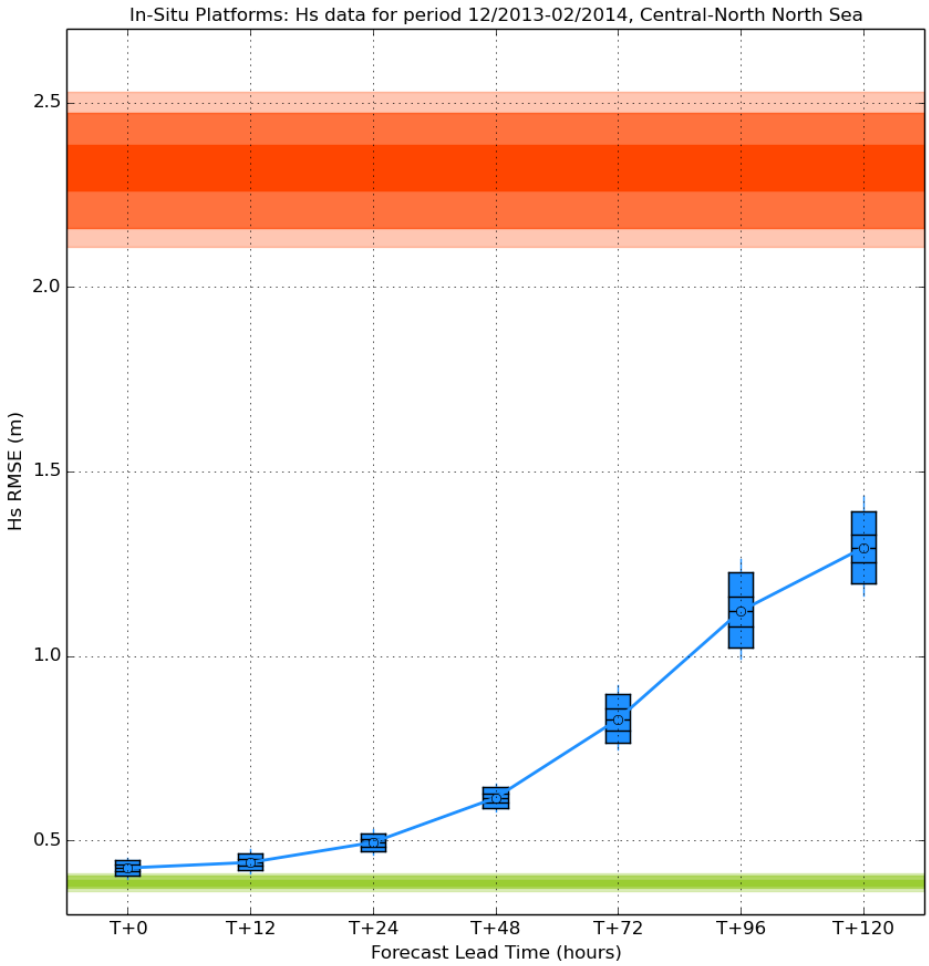


Verification: Root Mean Square Error (RMSE) plot

What does this tell me?

The plot works similarly to MAE and measures the range of errors occurring between observed and modelled instances of the given variable over the data sample. RMSE will increase with (absolute) bias and as the model's ability to replicate the temporal signal of the observations decreases and emphasises the contribution of large errors compared to the MAE.

In the example shown, RMSE at short lead times falls close to the idealised case, i.e. the majority contribution to model-observation errors is expected to be a result of observing error. At longer lead times the model errors are a number of times higher than the observed error. However, even at 5 days ahead (T+120) the model is significantly more skilful than a 'random guess'.

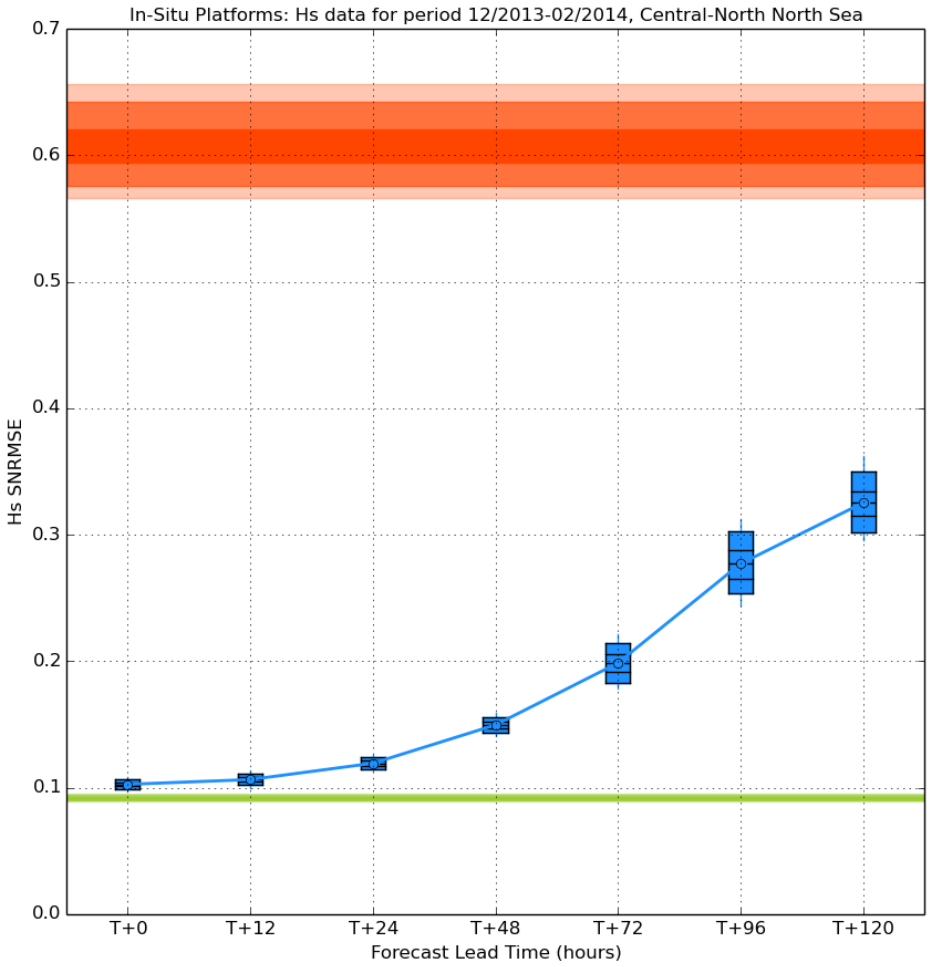


Verification: Symmetrically Normalised Root Mean Square Error (SNRMSE) plot

What does this tell me?

The plot contextualises RMSE against the background variability of the observations and model during the sample period. Percentages could be used instead of the normalised scale given in this example. SNRMSE will increase with (absolute) bias and as the model's ability to replicate the temporal signal of the observations decreases.

In the example shown, SNRMSE at short lead times falls close to the idealised case, i.e. the majority contribution to model-observation errors is expected to be a result of observing errors, which are of the order of 10% of the background variability in wave height. At longer lead times the model-observation errors are close to 30% of the range by which Hs will vary over the sample period. However, even at 5 days ahead (T+120) the model is significantly more skilful than a 'random guess'.

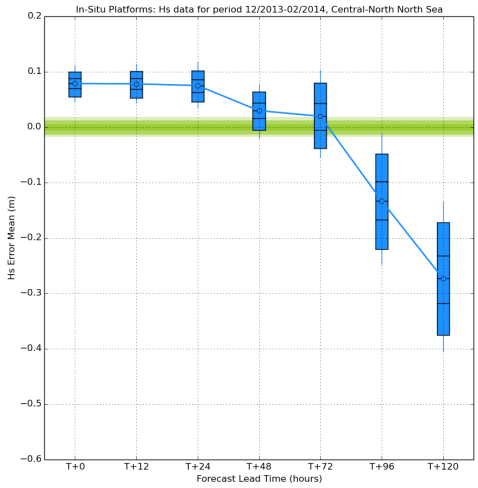


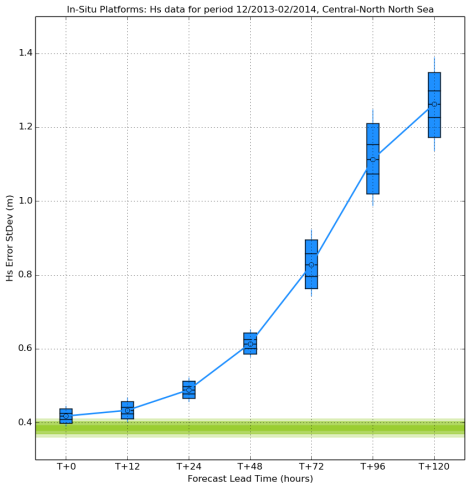
Test M2: Quantify features of the distribution of prediction errors

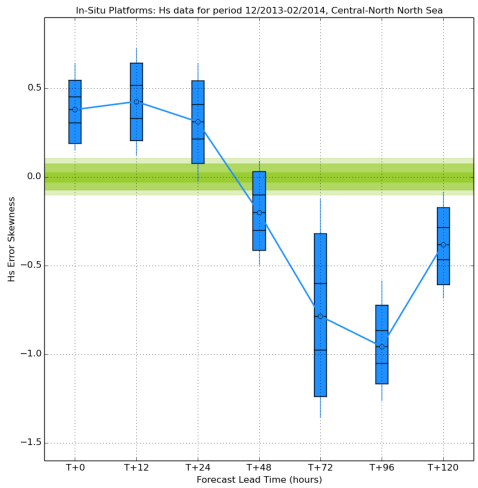
Verification: Error moments

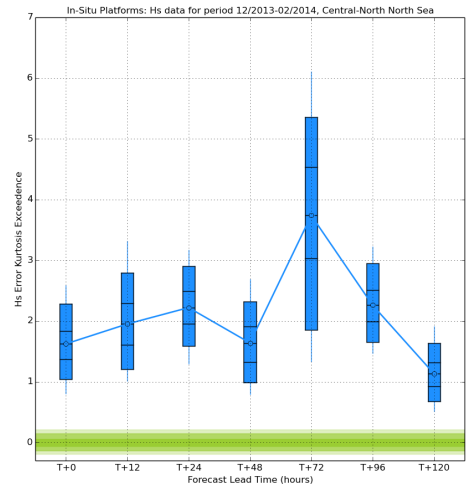
What does this tell me?

The presented '4-up' view shows forecast lead time changes in the model-observation error statistics representing 1st, 2nd, 3rd and 4th moments of the error distribution. These can be related to location, scale, skewness and kurtosis of a prescribed probability density function. The green plume shows the idealised moments. Good performance is identified when the plume and box-and-whiskers overlap for the first 3 moments (bias, standard deviation, skewness). In this example, performance at short lead times is reasonably good, although the model slightly overestimates conditions and positive skewness indicates a longer tail to overprediction errors than underprediction errors. The scale of errors increase with lead time and are accompanied by a negative bias (underprediction) and shift to negative skewness, which indicates that the largest errors will be associated with underprediction. The positive kurtosis exceedence statistic indicates that, compared to a normal distribution, a higher proportion of errors are clustered around the mean bias value, but that the tails of the distribution (associated with large errors) are also longer.









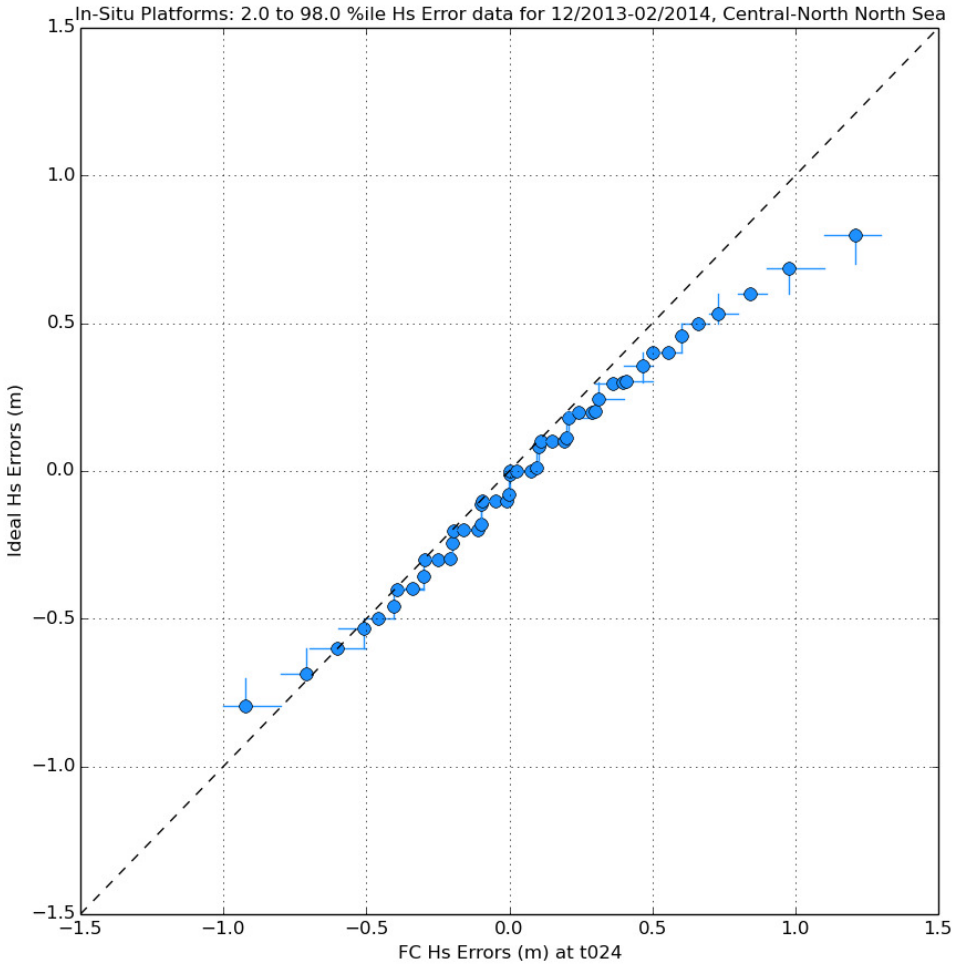
Test M3: Compare errors from two prediction systems

Verification: Error Q-Q plot

What does this tell me?

The plot compares frequency distributions of prediction-reference errors. Data points deviate from the 1:1 line when, for a given percentile of the frequency distribution, a higher value occurs in either data source. A constant offset of points from the 1:1 line will indicate a systematic difference in bias, whilst a curve away from the 1:1 line at lower or higher percentiles suggests that the error tail (larger absolute differences between model and observation) is longer in a particular data source.

In this example the data compared are T+24 model-observation errors and the idealised (observation only) errors. The data points are slightly offset from the 1:1 line, indicating a small bias toward the model-observation error data, and the curve away from the 1:1 line for the positive errors indicate that high overprediction errors (>0.5m) are significantly more likely for the model than would be expected if observations provided the only source of error.

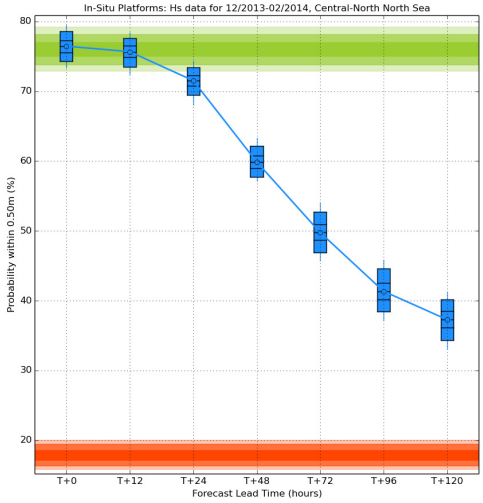


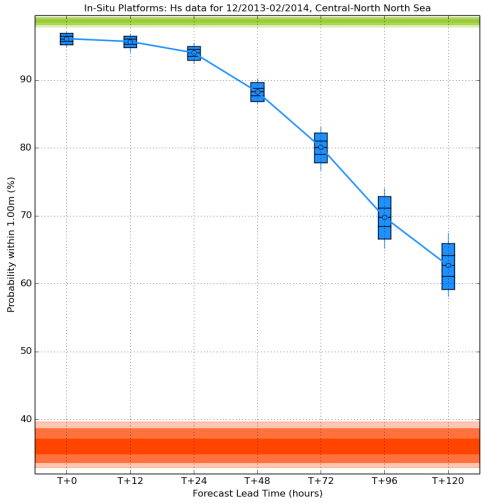
Test P1: Quantify likelihood of predictions to fall outside prescribed a tolerance

Verification: Probability within plot

What does this tell me?

The plot shows the (percentage) probability that a forecast falls within a predefined (absolute) range from the observed value. The statistic will become lower when model predictions do not successfully meet this criterion. For ranges set to a low value, this statistic can be used to define successful performance. For a range set to a high value the statistic defines the risk of a poor forecast being issued. Good performance occurs when the box-and-whiskers overlaps the idealised performance plume. The examples shown (respectively) test forecasts within 0.5m of the observation (success) and 1.0m (as a 'bust' threshold). Forecasts at short lead time are highly successful, but performance falls off significantly with increasing lead time. Bust probabilities are low out to T+24 (100 – y-axis value), but would be significant if a forecast with a lead time of T+48 or longer were used.





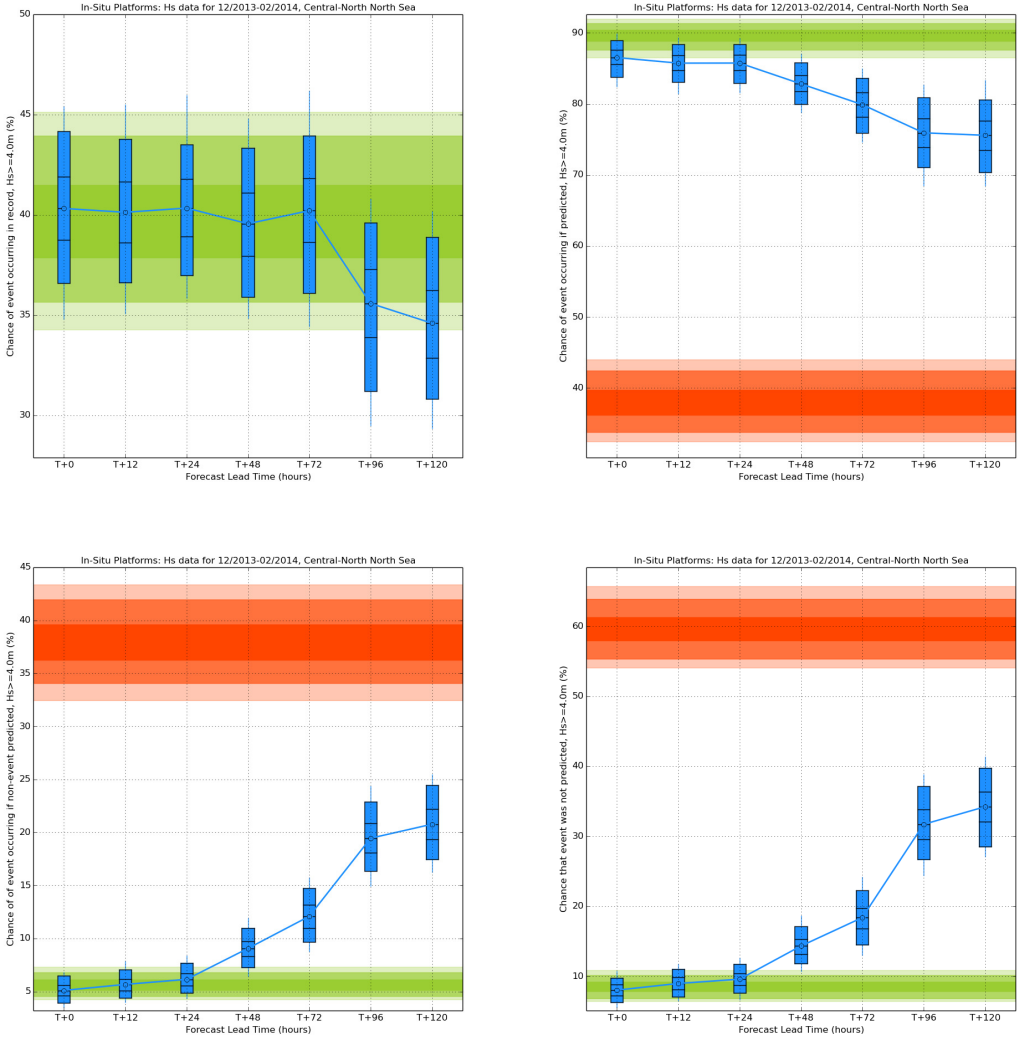
Test P2: Quantify ability to predict event x

Verification: Contingency table scores

What does this tell me?

The '4-up' plot shows (percentage) probability scores relating to forecast prediction of a given event. The top left plot compares the number of event occurrences in the forecast (box-and-whiskers) to the number of observations (plume). Good (climatological) performance occurs when the data overlap. The top right plot shows the chance that an event occurs if predicted (and the chance of the false alarm can be inferred as 100 – data value). The lower left plot indicates the chance that an event might occur when not predicted, and the lower right plot shows the probability that, given an event occurred, the event was not forecast. These statistics are highly dependent on climatology and the threshold being set, as indicated by the values in the orange plume.

In the case shown, performance of the forecast is close to optimal (considering the effect of observation errors as illustrated by the green plume) for the first forecast day, and the forecasts display skill out to day 5. However, the risks of missing an event incurred by following the forecast guidance directly at day 3 and beyond are significant.



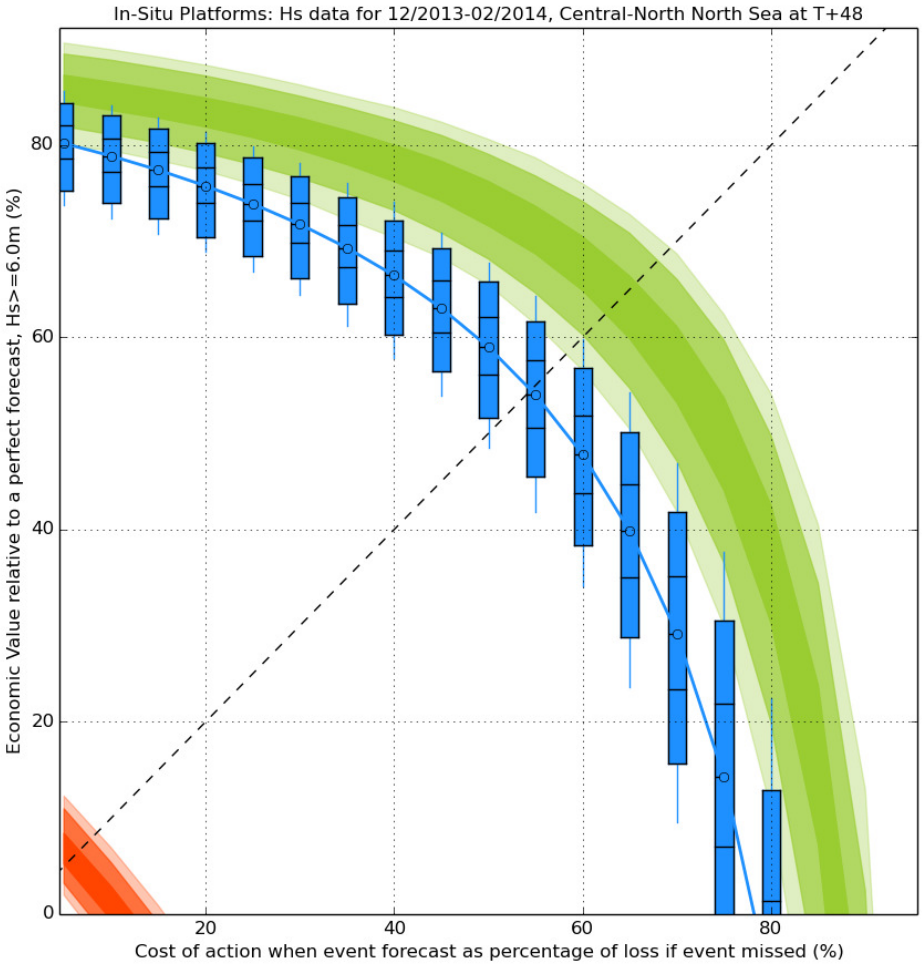
Test P3: Quantify long term benefit of decision making using predictions of event x

Verification: Relative Economic Value (REV) score

What does this tell me?

The plot shows the economic value of using the forecast model for event prediction compared to a 'perfect' system that predicts all events and non events correctly. The economic value is determined based on the relative cost of taking mitigating action versus the cost of missing an event, and is displayed on the x-axis. The dotted line denotes the value of a 'no forecast' system versus the perfect forecast, which increases as mitigating actions become more expensive. The forecast system performs well when the box-and-whiskers data overlap the green plume, and adds value when the data falls above the dotted line.

In the example given, the forecast at T+48 has appreciable skill and will add value over a 'no forecast' system so long as the cost of mitigation falls below 50% of the loss incurred when an event occurs without warning.



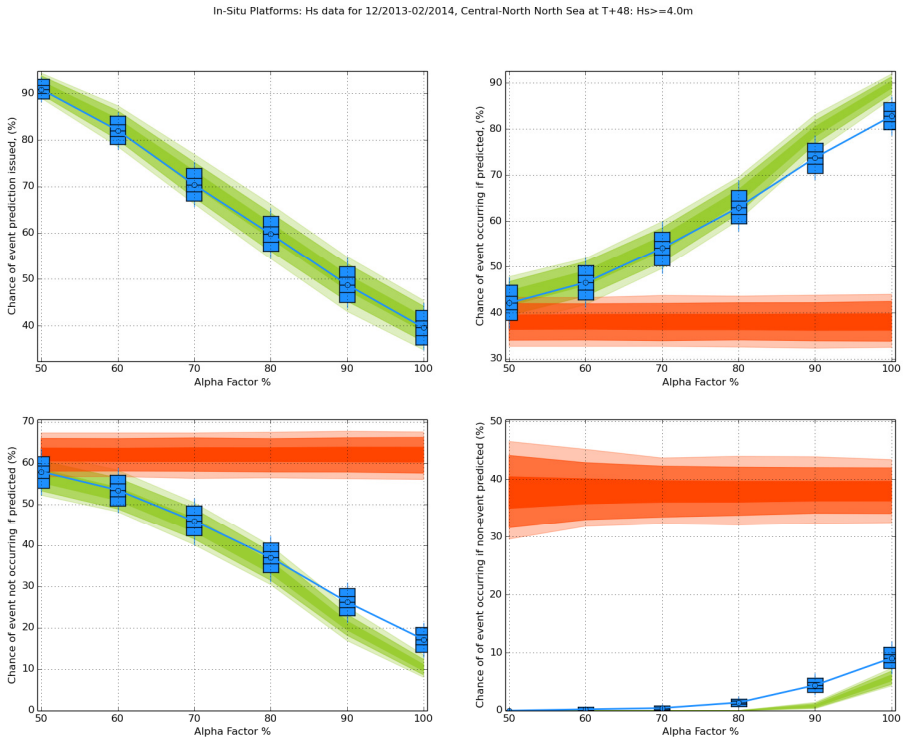
Test P4: Quantify effects of altering prediction threshold(s) for event x

Verification: Alpha factor contingency table scores

What does this tell me?

The '4-up' plot shows contingency tables scores (similar to P2) and their relative change for different 'alpha factor' values, i.e. when the operating threshold is met once the forecast achieves a value of operating threshold*alpha. The top left plot shows the number of event occurrences in the forecast. The top right plot shows the chance that an event occurs if predicted. The lower left plot indicates the chance of a false alarm (operation postponed) can and the lower right plot shows the probability that the event occurred when not forecast. These statistics are highly dependent on climatology and the threshold being set, as indicated by the values in the orange plume.

In the case shown, a user that did not have to be entirely risk averse would be able to use an alpha factor of 90-100% to minimise 'false downtime' whilst keeping the risk of a missed event at less than 10%. For a more risk averse user an alpha factor of 80% might be taken although the number of false downtime forecasts by a factor of 4.



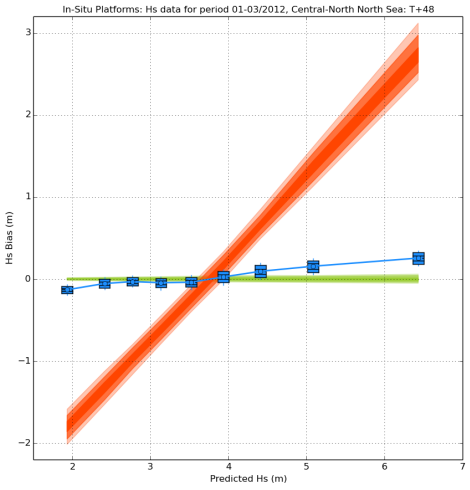
Test R1: Quantify errors through predicted event sub-ranges

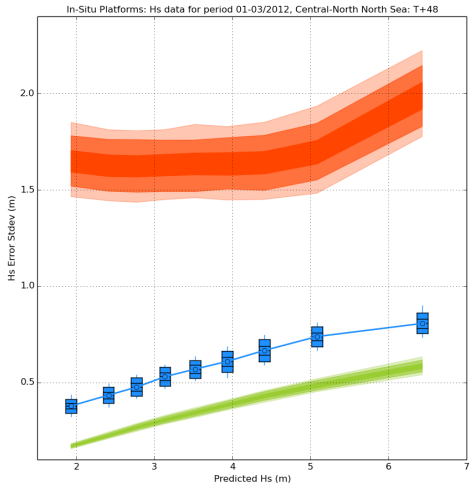
Verification: Error bias and standard deviation through predicted range

What does this tell me?

The plots show (respectively) the average error between modelled minus observed values of a given variable and standard deviation of the errors. Values are given in blocks representing partitions of the full data sample, based on the predicted value of the variable. A large bias (relative to background conditions) may be indicative of a fundamental difference between model predictions and the observations for a given prediction category. It should be noted that, if a model correctly predicts the range of conditions, but is not very highly correlated with observations, the lowest category might be expected to show a negative bias, and the highest category should be expected to show a positive bias. A large standard deviation (relative to predicted conditions) may be indicative of an inability of the model predictions to reflect the time signal of the observations. It should be noted that an increase in standard deviation is likely to be linked to an increase in predicted conditions for all but the most highly correlated predictions. Good performance occurs when the box-and-whiskers data match up with the green plumes.

In this example de-correlation of the model versus observations at T+48 means that the model is biased low for low wave height predictions and high for predictions above 4m. Model error standard deviation is significantly above that expected in the ideal case.



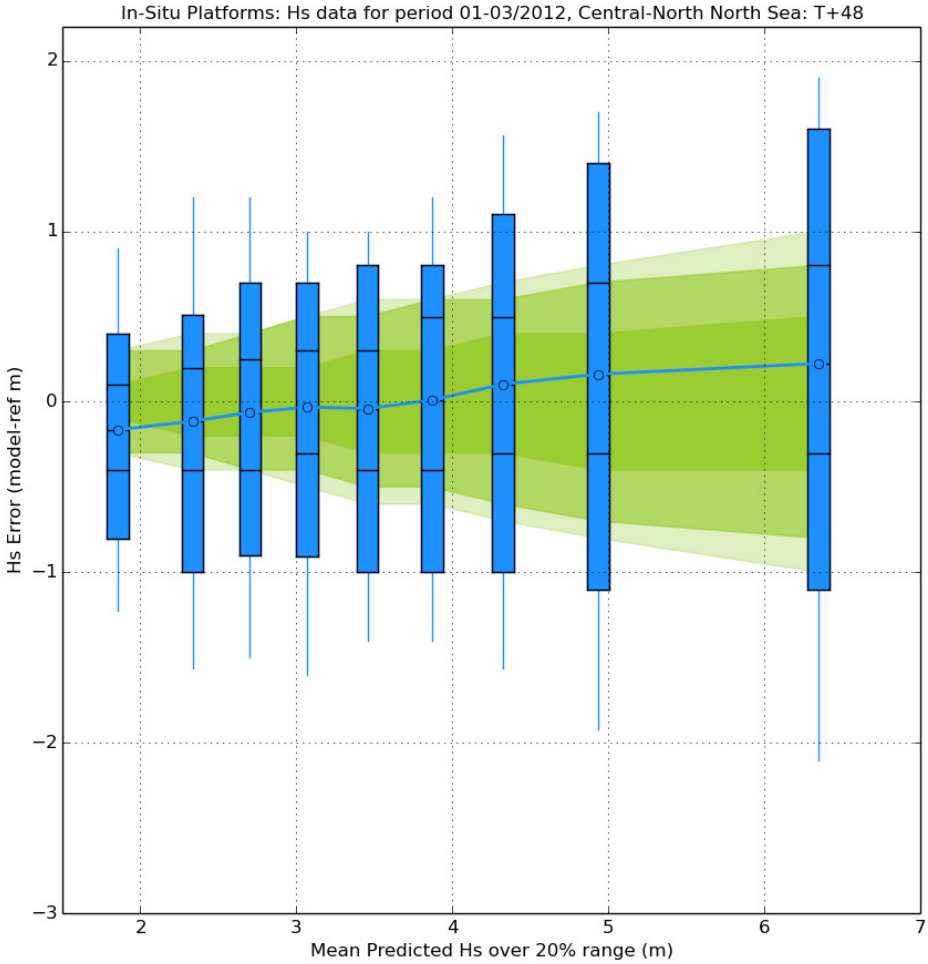


Verification: Error quantiles through predicted range

What does this tell me?

The plots show box-and-whiskers descriptions of model minus observation errors found based on the predicted value of the variable. Good performance occurs when the box-and-whiskers data match up with the green plumes. The centre circle defines the mean error value, the inner box lines define the inter-quartile range, the outer box lines define the 5-95% range and the whiskers define the 1-99% range for the errors.

This example uses the same dataset as for the bias and standard deviation plots shown previously and the whiskers illustrate skewness effects in the error distribution that cannot be seen in the other plots. In this case the data show that the model errors have significantly more spread than in the idealised case and that, for wave height predictions above approximately 4.5m a large overprediction is more likely than a large underprediction.



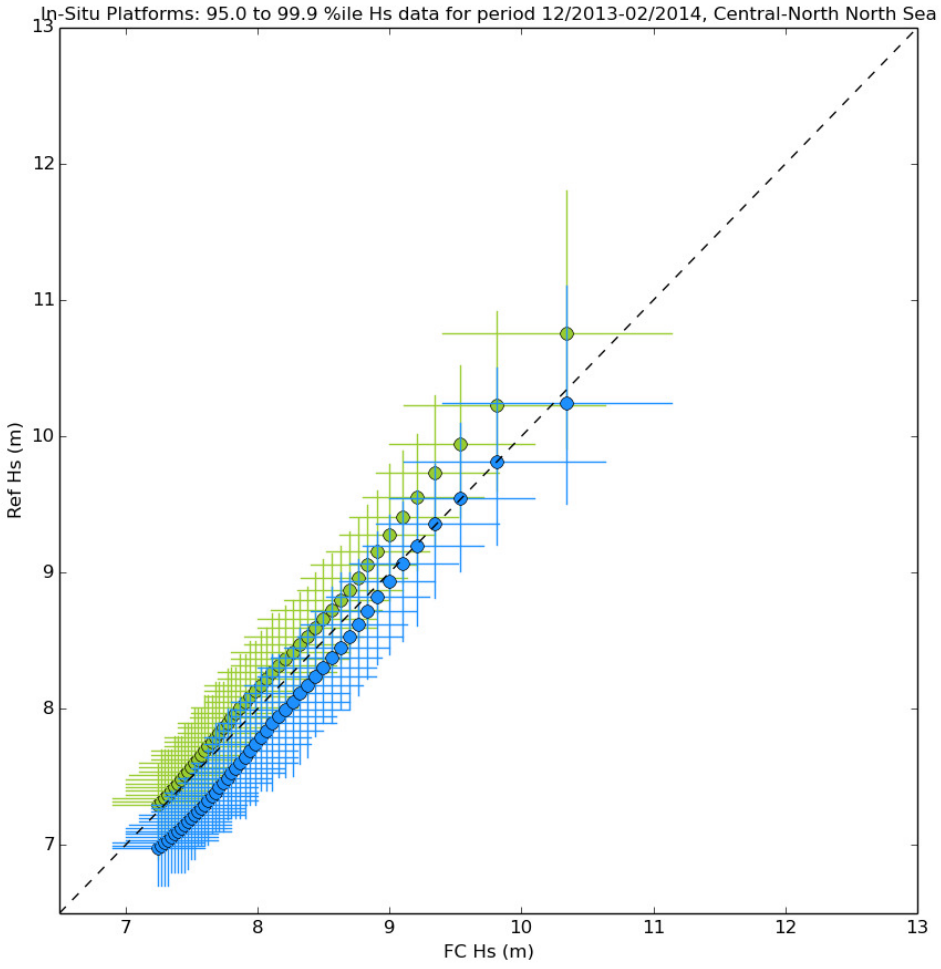
Test X1: Test that reference extremes are reproduced by the prediction system

Verification: Q-Q plot for upper percentiles

What does this tell me?

The plot compares frequency distributions of observed and modelled occurrence of the given variable. Data points deviate from the 1:1 line when, for a given percentile of the frequency distribution, a higher value occurs in either the observed or modelled data. For example a constant offset of points from the 1:1 line will indicate a systematic bias, whilst a curve away from the 1:1 line indicates that the distribution is under- or over-sampled.

In this example the plot compares percentiles above the 95% level. The location of the model data points (blue) compared to both the 1:1 line and location of the observed points (green) show a tendency to overpredict at higher percentiles. Note that the trend in the green symbols suggest that in an ideal scenario the model should underpredict high percentile observations – this is expected to be an effect of assuming observation errors are normally distributed around the wave height population tail.

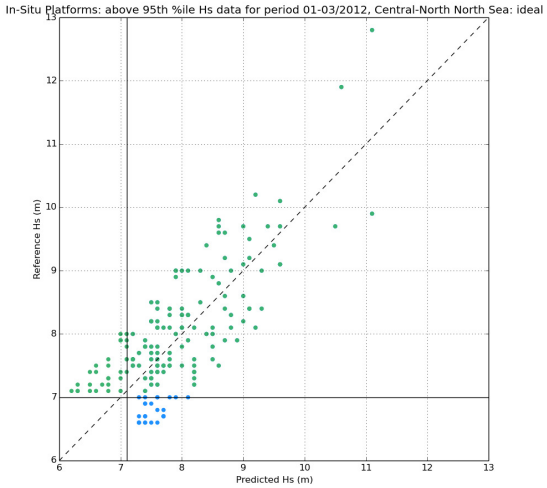


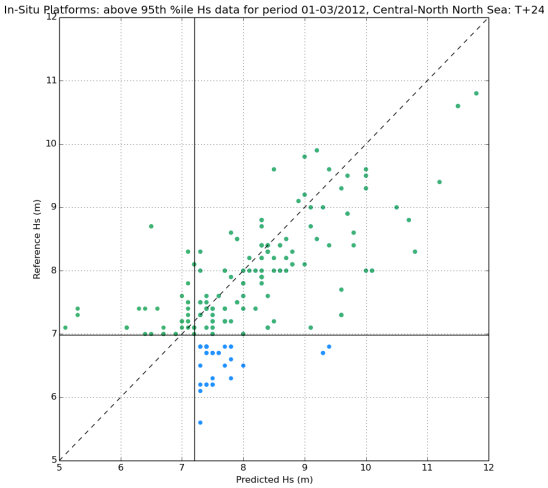
Test X2: Test that events in the tails of the predicted and reference distributions are well correlated

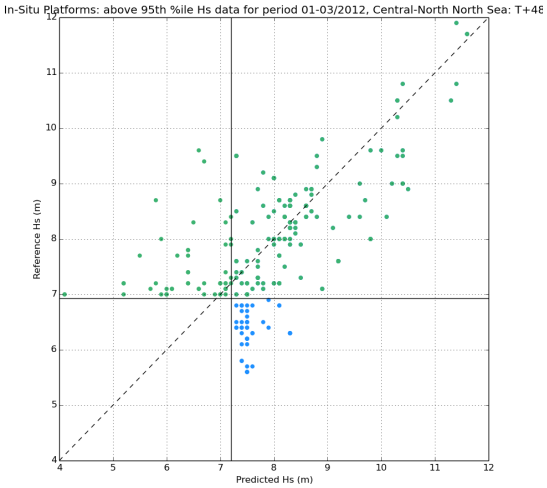
Verification: Scatter plot for upper percentiles

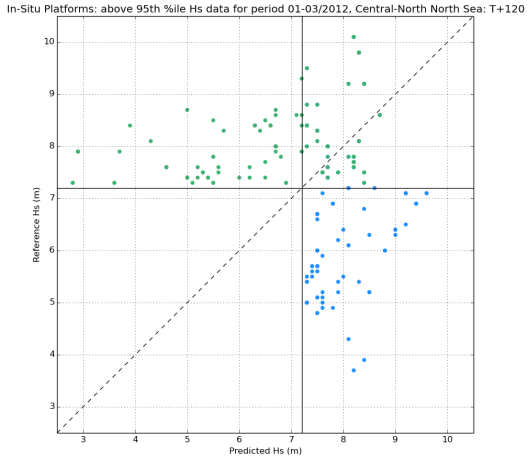
What does this tell me?

The plot compares scatter data generated based on the top 5% of predicted values and the top 5% of observed values. Data fall in the lower right quadrant of the plot when the prediction is below the observed 95th percentile and in the upper left quadrant of the plot when the observations are above the 95th percentile but the predictions aren't. In this '4-up' example the scatterplots at various lead times are compared with an idealised case based purely on observation errors. Event identification at T+24 is close to ideal, although the spread of errors is increased. Identification of events at T+48 is reasonable, but the value of the forecast at T+120 is questionable.









Map based statistics

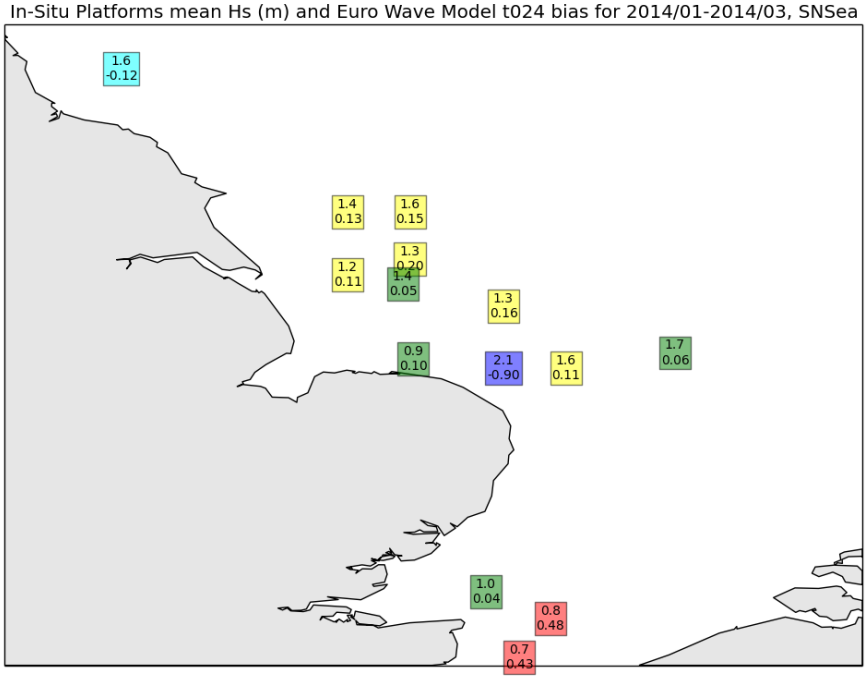
Map view: observed mean and model-observation bias/RMSE

What does this tell me?

These plots aim to show the geographical distribution of verification data for a given region. In the cases shown the data values are provided at the location of in-situ observing sites. The top value in the box is the mean observed value and the lower value the bias/rmse. Boxes are colour coded according to the value of the statistic. Contrasting both mean and statistic values between sites gives the user some feel for inconsistencies in either model prediction skill or observation behaviours.

For example, in the plots for the southern North Sea (shown overleaf) a general bias of order 0.1m is noted and an RMSE of 0.25-0.3m. Outliers are identified in the southern part of the area, but are where the observing platform is a lightvessel, which has known under-observation properties due to its large hull size. The outlier off the north coast of Norfolk is an oil industry platform that was only reporting sporadically through the verification period and may have had some quality control issues.

In-Situ Platforms mean Hs (m) and Euro Wave Model t024 bias for 2014/01-2014/03, SNSea



In-Situ Platforms mean Hs (m) and Euro Wave Model t024 RMSE for 2014/01-2014/03, SNSea