Consensus Forecasts of Modelled Wave Parameters

TOM H. DURRANT[1], FRANK WOODCOCK[2], AND DIANA J. M. GREENSLADE[3]
[1,2,3]*Centre for Australian Weather and Climate Research , Bureau of Meteorology*

ABSTRACT

The Operational Consensus Forecast scheme uses past performance to bias-correct and combine numerical forecasts to produce an improved forecast at locations where recent observations are available. This technique was applied to forecasts of Significant Wave Height (H_s), peak period and 10 m wind speed from 10 numerical wave models, at 14 buoy sites located around North America. Results show the best forecast is achieved with a performance-weighted consensus of bias-corrected components for both H_s and peak period, while a performance weighted composite of linear corrected components gives the best results for wind speed. For 24 hour forecasts, these composites produce improvements of 36%, 47% and 31% in RMSE values over the mean raw model components respectively, or 14%, 22% and 18% over the best individual model. Similar gains in forecast skill are retained out to five days. By reducing the number of models used in the construction of consensus forecasts, it was found that little forecast skill is gained beyond five or six model components, with independence of these components, as well as individual component quality being important considerations.

1. Introduction

The Operational Consensus Forecast (OCF) scheme of Woodcock and Engel (2005), combines forecasts derived from a multimodel ensemble to produce an improved real-time forecast at locations where recent observations are available. Component model biases and weighting factors are derived from a training period of the previous 30 days of model forecasts and verifying observations. The next real-time OCF forecast is a weighted average of the set of latest-available, bias-corrected, component forecasts. Each component forecast is weighted by the inverse of the mean-absolute-error (MAE) of that forecast over the training period. In operational daily weather prediction at the the Australian Bureau of Meteorology (the Bureau), OCF combines both operationally available model output statistics forecasts (MOS; Glahn and Lowry 1972) and bilinearly interpolated direct model output forecasts at over 700 sites twice daily from 0 to 7 days ahead. OCF superseded MOS as the official objective forecast guidance in March 2005.

Recently, Woodcock and Greenslade (2007) investigated the application of OCF techniques to wave forecasts. They employed OCF to generate 24 hour predictions of Significant Wave Height (H_s) at 18 observation locations around Australia. Broad conclusions were that in deep water, a 20% - 30% improvement over model forecasts of H_s can be achieved using the OCF strategies and in shallow water, the strategy of compositing model forecasts after linear correction can yield a 60% - 70% improvement over raw model forecasts. However, this work was hampered by a lack of quality independent models for compositing, with only five models available, two of which were high resolution nested models within a third, resulting in a lack of independence between these three.

This is addressed in this work with ten independent models from the major forecasting centres used for compositing. Direct Model Output (DMO) forecasts, interpolated from these models to 14 moored buoy sites surrounding North America provide the underlying component forecasts in the OCF composite. In addition to H_s, the application of OCF techniques to both wind speed and peak period at these same sites is also investigated, with the analysis extended to cover increased forecast periods out to five days. The question of the dependence of the performance of OCF schemes on the number of component models used is also addressed more directly by looking at the effects of reducing the number of component models.

Both model and observational data are examined in section 2, a general description of the OCF techniques and the specific application used here are described in section 3, results are presented in section 4 and finally

Corresponding author address:
T. H. Durrant,
Centre for Australian Weather and Climate Research
Bureau of Meteorology
GPO Box 1289
Melbourne, VIC 3001 Australia
E-mail: t.durrant@bom.gov.au

section 5 contains a summary of the work.

2. Data

For the past six years, a monthly exchange of ocean wave model data has been taking place between the major forecasting centres around the world (Bidlot et al. 2002). What started as a cooperation between the European Centre for Medium-Range Weather Forecasts (ECMWF), the The U.K. Met Office (UKMO), Fleet Numerical Meteorology and Oceanography Center (FNMOC), the Meteorological Service of Canada (MSC) and the National Centers for Environmental Prediction (NCEP) has now grown to include in chronological order of participation, Deutscher Wetherdienst (DWD), the Bureau, Meteo-France (METFR), the French Hydrographic and Oceanographic Service (SHOM), the Japan Meteorological Agency (JMA), the Korean Meteorological Administration (KMA) and the Puertos del Estados (PRTOS). On a monthly basis, each centre provides files of model analysis and forecast data to the ECMWF at an agreed list of moored buoy sites at which instrumented observations of H_s, wave period and wind speed are available, with results then sent out to participating centres. It is this data set that provides the basis for this work.

a. Observational Data

Observational data comes from moored buoys. Buoy data are generally assumed to be of high quality, and have been used in numerous studies for validation of model (eg Janssen et al. 1997; Caires and Sterl 2003; Caires et al. 2004) and altimeter (eg Tolman 2002; Queffeulou 2004; Faugere et al. 2006) data. As part of the collating process performed at ECMWF, these data undergo a quality control process to remove suspect observations. Wind speeds are adjusted to 10m height, and spatial and temporal scales are made comparable by averaging the hourly observations in time windows of four hours centered on the synoptic times. Full details of this process can be found in Bidlot and Holt (2006).

Over the course of the project, the number of model outputs available at each buoy has increased, as have the number of validation sites as new participants contribute additional buoy data from their respective institutions. The full list of buoys now includes some 245 locations. However many of these buoys are recent additions, and contain short time series of historical data. Others have only a subset of model data available at the site.

In order to achieve a clean data set with the maximum number of models, a subset of buoys was chosen here for which all participating models were present. KMA and PRTOS joined the intercomparison only recently (July 2007) and were not used in this work. SHOM and JMA joined in October 2006, and the desire to include these models determined the period examined from October 2006 through July 2007. These buoys are shown in Figure 1, with detail of each buoy presented in Table 1. All these buoys are classified as deep water buoys, well exceeding the depth limitations of operational global wave models, which typically provide wave forecasts that are skillful only in water depths greater than about 25m (Booij et al. 1999).

Table 2. Wave model characteristics including the domain over which the model is run, the grid resolution, the classification of the model and whether or not the model includes data assimilation. A mixed domain refers to the case where regional models are nested within a global model in which case the grid resolution refers to that of the global model.

Model	Domain	Grid. Res	Classification	DA
ECMWF	Mixed	0.36°	WAM	Yes
UKMO	Mixed	$\frac{5}{6} \times \frac{5}{9}^\circ$	Second Gen	Yes
FNMOC	Global	0.5°	WW3	No
MSC	Regional	1.0°	WAM	Yes
DWD	Mixed	0.75°	WAM	No
NCEP	Mixed	$1.25 \times 1^\circ$	WW3	Yes
AUSBM	Global	1.0°	WAM	Yes
METFR	Global	1.0°	Second Gen	Yes
SHOM	Mixed	1.0°	WW3	No
JMA	Mixed	1.25°	MRI-III	No

The buoys used here are operated by either the National Data Buoy Center (NDBC) or the Marine Environmental Data Service (MEDS). Recent work (Durrant et al. Submitted) suggests that systematic differences may exist between these two networks, with MEDS reported H_s values being 10% low relative to those reported by NDBC buoys. This is of limited relevance here, as each site is treated independently. It is worth noting however, that OCF techniques are limited by the accuracy of the observations available.

b. Model Data

Some details of the models used in this work are provided in Table 2. The provision of full details of all these models is impractical here, further references can be found in Bidlot et al. (2007). Two models dominate this list, namely the third generation WAve Model (WAM; WAMDI-Group 1988; Komen 1994) and WAVEWATCH III (WW3) (Tolman 1991). Operational versions of these models have, however, undergone many independent changes and tunings. All models also have different wind forcing, spatial resolutions, data assimilation systems etc. These differences result in errors that vary between the models, thereby enhancing the potential gain from a consensus forecast.

3. OCF Methodology

a. Description

The OCF methodology of Woodcock and Engel (2005) is a simple statistical scheme, which takes a weighted average of bias - corrected component model forecasts on

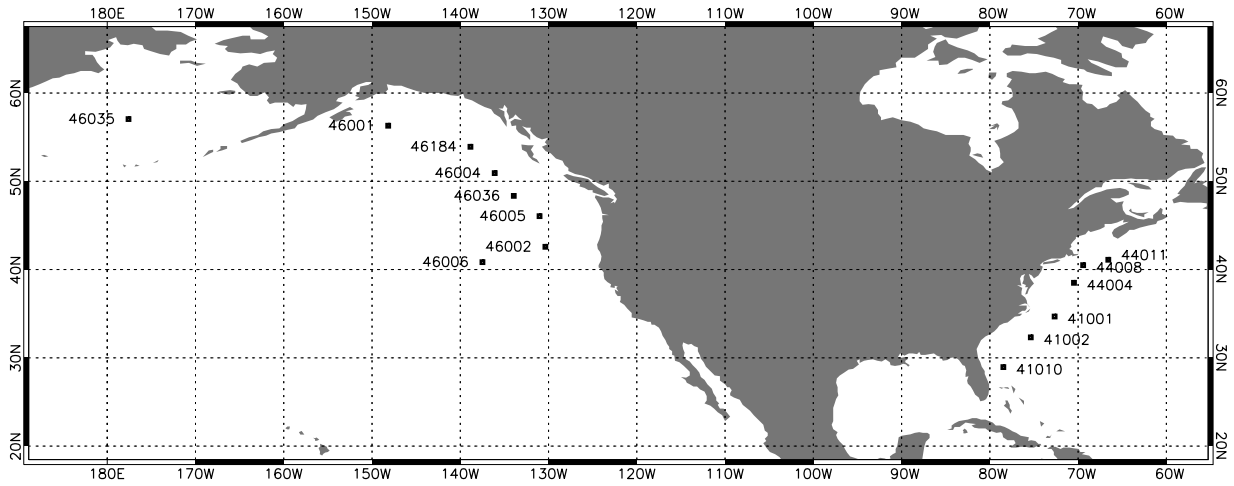

FIG. 1. Location of buoys used in this study (see table 1).

Table 1. Details of buoys used in this study. DMO refers to the number assigned by the World Meteorology Organization

WMO	Name	Owner	Latitude	Longitude	Depth (m)
41001	E Hatteras	NDBC	34.68	-72.66	4427
41002	S Hatteras	NDBC	32.32	-75.36	3316
41010	Cape Canaveral East	NDBC	28.95	-78.47	873
44004	Hotel	NDBC	38.50	-70.47	3182
44008	Nantucket	NDBC	40.50	-69.43	62
44011	Georges Bank	NDBC	41.11	-66.58	88
46001	Gulf of Alaska	NDBC	56.30	-148.17	4206
46002	Oregon	NDBC	42.58	-130.36	3525
46004	Middle Nomad	MEDS	50.93	-136.10	3737
46005	W Astoria	MEDS	46.05	-131.02	2780
46006	SW Astoria	NDBC	40.84	-137.49	4023
46035	Bering Sea	NDBC	57.05	-177.59	3658
46036	South Nomad	MEDS	48.35	-133.94	3676
46184	North Nomad	MEDS	53.91	-138.85	3406

a site-by-site and day-by-day basis. The scheme is based upon the premise that each model derived forecast (f_i) has three components: the true value (o), a systematic error component or bias (b_i) that can be approximated and removed, and a random error component (e_i) that can be minimised through compositing (i indicating each separate model). The success of the OCF scheme is based upon the estimation of bias and weighting parameters.

Bias and weighting parameters are based on a moving window of historical data. Model biases (b_i) are approximated using the best easy systematic estimator (BES; Wonnacott and Wonnacott 1972, section 7.3) over the errors in the sample:

$$\hat{b}_i = BES = \frac{(Q_1 + 2Q_2 + Q_3)}{4} \tag{1}$$

where Q_1, Q_2 and Q_3 are the error sample first, second, and third quartiles, respectively. This is more robust than a simple arithmetic mean. Normalised weighting parameters ($\hat{w}_i$) are calculated by using the inverse MAE from the bias-corrected error samples of the n contributing model forecasts over the training period, with

$$\hat{w}_i = (MAE)_i^{-1} \left(\sum_{i=1}^{n} (MAE)_i^{-1} \right) \tag{2}$$

Using these parameters, OCF based on n model forecasts (f_i) is given by:

$$OCF = \sum_{i=1}^{n} \left(\hat{w}_i \left[f_i - \hat{b}_i \right] \right) \tag{3}$$

Breaking the forecasts (f_i) into the aforementioned com-

ponents,

$$OCF = \sum_{i=1}^{n} \left(\hat{w}_i \left[(o + b_i + e_i) - \hat{b}_i \right] \right) \quad (4)$$

Gathering terms this becomes:

$$OCF = o + \sum_{i=1}^{n} \left(\hat{w}_i \left[b_i - \hat{b}_i \right] \right) + \sum_{i=1}^{n} (\hat{w}_i e_i) \quad (5)$$

The final two terms in equation 5 highlight the importance of the bias removal and weighting schemes. Characterisation of the random nature of the error distributions, as part of the weighting scheme, aids minimisation of the random errors via compositing with highly variable models penalised for their reduced reliability.

b. Application

As in Woodcock and Greenslade (2007) a number of corrected forecasting techniques based around this technique are explored. Internal methods are those in which the model forecast is corrected according to a training set based on that particular model. Two internal methods were used to modify the direct model output forecasts. The first was a simple bias correction using BES. The second was a least squares linear-regression correction whereby a linear-regression equation between the predictands (observations) and predicators (direct model outputs) was generated and then applied to the next forecast.

Several forms of compositing were investigated. The simplest is the average of all components, referred to here as equal weighting since the components are equally weighted. Performance weighting combines the forecasts according to their performance over a training set, as described above. Both equal weighted and performance weighted forecasts were produced from both bias corrected and linearly corrected model components (referred to as equal weighted bias correction (EWBC), equal weighted linear correction (EWLC), performance weighted bias correction (PWBC), and performance weighted linear correction (PWLC))

Finally, we generated forecasts by using the linear-regression coefficient and intercept derived from the best-performing linear-regression-corrected component in the training period at a site and applying them to the corresponding next independent component forecast for that site (i.e., the coefficient, intercept values and component model change for every forecast). This is referred to as the best linear corrected (BLC) forecast. All bias-correction, linear-regression-correction and BLC comparisons are undertaken over corresponding, matching events (i.e., identical verifying sets and training windows).

The effect of varying the training period was investigated by Woodcock and Greenslade (2007) by increasing the training window in steps of four from 1 to 59 events. They found that bias corrected methods stabilised by 9 events, and linearly corrected methods by 13 events. In order to maintain consistency with the bulk of this work, a fixed training window of 29 events was used here.

4. Results

We begin by examining results for 24 hour H_s forecasts. The same correction techniques are then extended to 24 hour forecasts of wind speed and peak period. The performance of composite forecasts is then examined over longer forecast periods and finally, variations in the number of component models included in the consensus forecasts are explored.

It is not the intention here to examine in any depth the performance of individual models, but rather to focus on the performance of the various composite schemes. Analysis of model performance based on this data set can be found in Bidlot et al. (2002) and more recently Bidlot et al. (2007). While these studies consider results in various regions, all buoys are considered together here. For the purpose of intercomparison and model diagnostics, the separation serves the valuable purpose of providing further insight into sources of error by comparing areas of wind sea or swell dominated areas for example, or areas where various sheltering and sub grid scale processes are of differing importance. However for this work, while the performance of the statistical scheme will differ with the quality of the input models, due to its non-physical nature, little is gained by examining regions separately. The same is also true for examining seasonal variation in error.

Verification statistics include bias, MAE, root-mean-square-error (RMSE), maximum absolute error (XAE), scatter index (SI = standard deviation normalised by the observation mean) and the percentage of explained variance (V% = $100\times$ the square of the correlation between forecast and observation). Statistics are calculated for each buoy, and overall statistics are calculated by averaging these results, weighted according to the number of observations at each buoy.

a. 24 hour forecast results

Figure 2 shows scatter plots for buoy and model colocations for each raw model, as well as BLC and PWBC results. Corresponding statistics for these plots are presented in Table 3 . ECMWF significantly outperformed the other models over this time period and set of buoys, yielding the lowest MAE, RMSE, XAE, SI, and the highest %V. This model also achieved a negligible overall bias, suggesting that little will be gained by bias correction. The superior performance of this model extends to individual buoys, being the best performer at 8 of the 13 buoys used here. All results for raw models, bias corrected and linearly corrected models, as well as the vari-

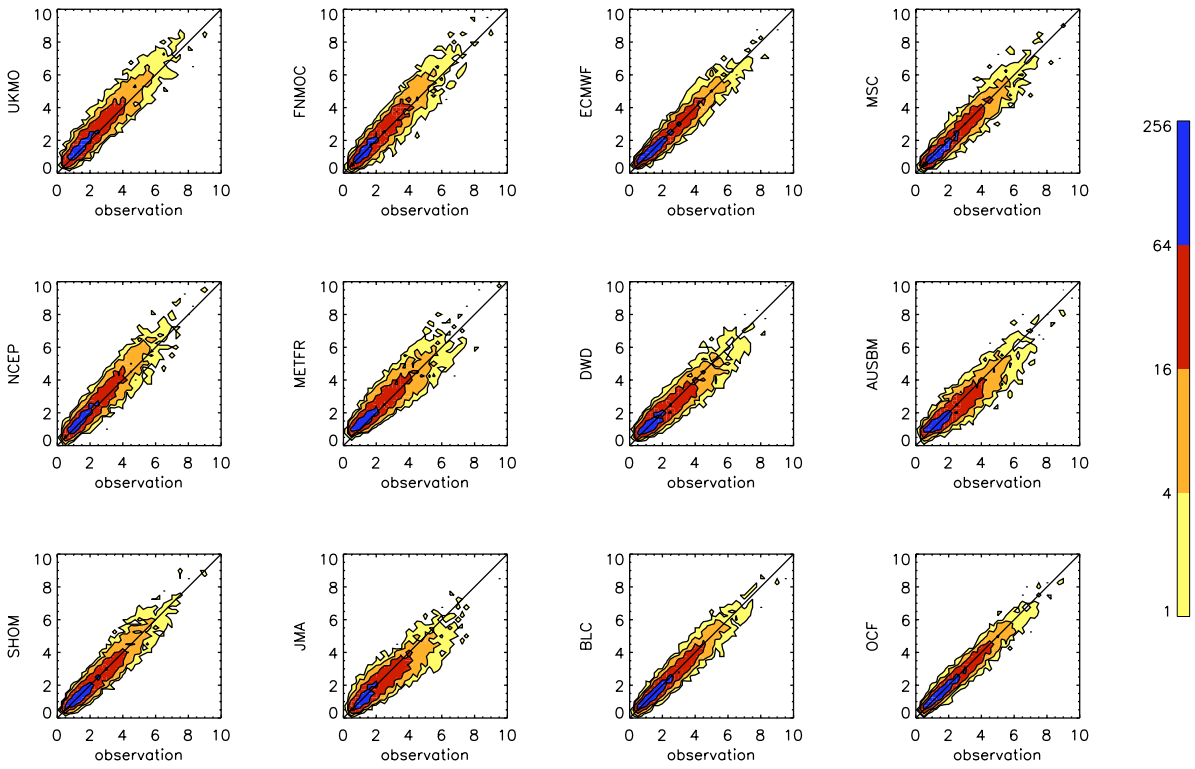

FIG. 2. Scatter plots of 24 hour forecast H_s for all raw models and BLC and PWBC (OCF) correction schemes

Table 4. Percentage improvement in RMSE for each model (see table 3) due to learned bias and linear correction schemes.

Model	Bias Correction	Linear Correction
UKMO	12	15
FNMOC	12	13
ECMWF	5	5
MSC	7	5
NCEP	14	16
METFR	10	10
DWD	12	8
AUSBM	7	5
SHOM	7	7
JMA	8	7

ous composite schemes can be found in Appendix 1.

Table 4 shows the improvement for individual models from both bias correction and linear correction schemes. While some models show significant improvements such as FNMOC, NCEP, METFR, and DWD, overall, gains are modest. The best correction scheme varies amongst the models, with no method clearly better than other.

Table 5 shows the improvements of various correction schemes over the average raw model error. As evident from table 4, on average, bias correction and linear correction show similar impacts of around 10% improvement. The "best" model (defined here as that with the lowest RMSE) shows significant improvement over the average. In this case, ECMWF is the best raw model, as well as the best bias and linearly corrected model. As seen in table 4, little improvement is made on this model by applying these corrections, resulting in little difference between the numbers seen here. Of the composite schemes, bias corrected schemes are outperforming linearly corrected schemes, and likewise performance weighted schemes come out better than equal weighted schemes. The best performer by all measures is the PWBC, producing a substantial improvement of 36% in RMSE over the average raw model error.

This is an encouraging result, indicating that despite the dominance of a single model, a performance weighted composite is able to beat it. This addresses one of the questions raised by Woodcock and Greenslade (2007), which found that the best performing model in that case was hard to beat with a consensus forecast. The results here suggest that this was due to a lack of models included in the composite.

Similar to table 3 for H_s, table 6 shows raw model statistics for wind speed. Full wind speed results for

Table 3. H_s statistics for all raw models used in this study.

MODEL	N	BIAS	MAE	RMSE	XAE	SI	V%
UKMO	4600	0.21	0.40	0.52	2.07	0.18	88.50
FNMOC	4600	0.09	0.39	0.52	2.29	0.19	88.36
ECMWF	4600	**-0.01**	**0.27**	**0.38**	**1.70**	**0.15**	**92.18**
MSC	4600	-0.09	0.32	0.44	2.31	0.17	89.85
NCEP	4600	0.16	0.38	0.51	2.07	0.18	88.94
METFR	4600	0.23	0.46	0.60	2.45	0.22	83.66
DWD	4600	-0.07	0.38	0.51	2.22	0.19	87.41
AUSBM	4600	-0.11	0.42	0.57	2.53	0.23	82.89
SHOM	4600	**-0.01**	0.33	0.44	2.09	0.17	90.19
JMA	4600	-0.15	0.44	0.59	2.50	0.23	82.21

Table 5. Percentage improvements over average of raw model error. Average BC and LC refer to the average corrected model error for each scheme, best BC and best LC refer to the best hindsight model after correction (based on RMSE).

Model	MAE (m)	RMSE (m)	XAE	SI	V%
Ave. raw	0.38	0.51	2.22	0.19	87.42
		Improvement over average of raw models (%)			
Ave. BC	11	9	4	1	-0
Ave. LC	12	9	-2	-0	-1
Best raw	29	26	24	21	5
Best BC	31	29	24	21	5
Best LC	31	28	22	20	5
PWBC	**38**	**36**	**35**	**29**	**7**
EWBC	35	34	32	27	7
PWLC	35	33	30	25	6
EWLC	32	30	25	22	6
BLC	28	25	12	17	4

each raw, bias corrected, linearly corrected, and composite schemes can be found in Appendix B. It should be noted that SHOM uses ECMWF winds though at 0.5°spatial and 6 hourly temporal resolution rather than the 40 km, 15 minute resolutions used operationally at ECMWF (Ardhuin, Personal Communication, 2006). For this reason, statistics for these winds are very similar, however, surprisingly, MAE and RMSE for SHOM winds are in fact lower for these buoys than that for ECMWF. This, it seems, is due to an increased positive bias in the ECMWF winds, with SI being the same for both centres. Bias corrected results yield little difference between the two in terms of MAE and RMSE, while ECMWF comes out marginally ahead under linear correction.

All models show a positive bias. Examining each buoy individually shows that this positive bias is present on the east coast only, however it is beyond the scope of this work to suggest why this bias exists. Of more relevance here, is that the presence of this bias suggest that a learned bias correction may have a positive impact on model performance.

Percentage improvements of the various correction schemes relative to the average raw model error are shown in Table 7. Unlike H_s which showed similar improvements for bias and linear correction schemes, wind speed performs better under a linear correction with more than double the improvement in the average linear corrected model RMSE (16%) relative to the average bias corrected model RMSE (6%). As with H_s, performance weighted composites outperform equal weighted composites, and as might be expected from the corrected models results, the PWLC composite slightly outperforms the PWBC composite.

Raw model statistics for peak period are shown in Table 8, again with full results for each raw, bias corrected, linearly corrected, and composite schemes to be found in Appendix C. Once again, the ECMWF model is the best model here, with the best MAE, RMSE, SI and %V values. The %V values are typically far lower than those seen for H_s and wind speed. This reflects the difficulties associated with the verification of this variable. Peak period refers to the period corresponding to the peak of the wave spectrum. As such, slight errors in the spectral shape can lead to large errors in peak period values. For example, in the case of a bimodal spectrum with two near equal peaks corresponding to wind sea and swell components, small errors in the energy associated with either part of the spectrum can lead to a large jump in the peak period as it moves from one peak to the other.

Percentage improvements coming from the same cor-

Table 6. Same as Table 3 for wind speed

MODEL	N	BIAS	MAE	RMSE	XAE	SI	V%
UKMO	4291	0.81	1.52	1.98	9.42	0.23	80.36
FNMOC	4291	0.55	1.64	2.14	9.79	0.26	73.12
ECMWF	4291	0.46	1.29	1.71	8.67	**0.21**	**81.24**
MSC	4291	0.49	1.52	2.02	8.68	0.25	74.41
NCEP	4291	0.71	1.65	2.15	**8.62**	0.26	74.52
METFR	4291	**0.16**	1.44	1.91	9.16	0.24	75.90
DWD	4291	0.18	1.46	1.92	8.56	0.24	73.91
AUSBM	4291	0.71	1.86	2.40	10.30	0.29	66.78
SHOM	4291	0.38	**1.28**	**1.69**	8.73	**0.21**	81.04
JMA	4291	0.68	1.63	2.12	9.31	0.25	75.79

Table 7. Same as table 5 for wind speed.

Model	MAE (m)	RMSE (m)	XAE	SI	V%
Ave. raw	1.53	2.00	9.12	0.25	75.71
Improvement over average of raw models (%)					
Ave. BC	8	6	-0	-1	-1
Ave. LC	16	16	13	10	-1
Best raw	16	16	4	14	7
Best BC	22	20	5	13	7
Best LC	27	25	16	19	6
PWBC	29	29	26	23	**11**
EWBC	28	28	28	22	**11**
PWLC	**31**	**31**	31	**26**	10
EWLC	29	29	**32**	24	9
BLC	25	23	16	18	5

Table 8. Same as table 3 for peak period.

MODEL	N	BIAS	MAE	RMSE	XAE	SI	V%
UKMO	4259	0.98	2.10	2.91	10.74	0.28	17.57
FNMOC	4259	-0.79	1.47	2.10	10.72	0.20	41.38
ECMWF	4259	0.34	**0.97**	**1.64**	10.21	**0.17**	**56.66**
MSC	4259	**-0.01**	1.05	1.73	9.66	0.19	48.53
NCEP	4259	-0.92	1.31	1.90	9.21	0.18	49.07
METFR	4259	-0.72	1.46	1.94	**7.57**	0.19	38.50
DWD	4259	-3.66	3.81	4.64	12.69	0.30	4.12
AUSBM	4259	-0.21	1.79	2.68	13.15	0.28	17.45
SHOM	4259	0.61	1.31	2.26	11.97	0.23	35.48
JMA	4259	-1.54	2.00	2.58	9.51	0.22	23.78

rection schemes for peak period are shown in Table 9. Individual models seem to respond well to learned correction schemes, with bias correction and linear correction resulting in an average 18% and 27% improvement in RMSE over the average raw model error respectively. Once again, performance weighted composites outperform equal weighted composites, and bias corrected composites outperform linearly corrected composites. BLC also performs well, giving a 37% improvement over the average raw model RMSE. This is, however, likely due to the high quality of the best model compared to the average raw model, which would be expected to feature heavily in the BLC forecast.

In the case of each variable (H_s, wind speed and peak period), results have been presented here in terms of improvement over the average raw model errors. While this gives an indication of what can be done with compositing techniques, and the improvements that can be gained over a set of input models, this kind of relative error gives a limited picture of the actual gains being achieved. In the case of peak period for example, PWBC achieves an impressive 47% improvement over the average raw model

Table 9. Same as Table 5 for wind speed

Model	MAE (m)	RMSE (m)	XAE	SI	V%
Ave. raw	1.73	2.44	10.54	0.22	33.25
		Improvement over average of raw models (%)			
Ave. BC	22	18	6	4	9
Ave. LC	27	27	19	14	11
Best raw	44	33	3	23	70
Best BC	44	34	5	23	70
Best LC	42	38	16	27	65
PWBC	**49**	**47**	34	**38**	**101**
EWBC	44	44	**39**	34	91
PWLC	42	42	31	31	83
EWLC	37	38	33	27	72
BLC	43	37	12	25	61

RMSE. However, the best raw model is 33% better than the average indicating a large spread in the quality of the models with respect to this parameter. Hence, it is important to consider not only gains over the average component models, but also gains over the best individual component.

Table 10 shows for each variable, the best performing corrections scheme, the improvement over the average error of the raw components and the improvement over the best raw model with respect to RMSE. H_s and peak period see far greater gains relative to the average error than those relative to the best model, while this is less so for wind speed. These results reflect the relative spread in the quality of the modelled variable, with more consistency seen in the modelled wind fields across the various institutions than the wave model variables.

Table 10. Gives the best performing correction scheme for each variable, and the RMSE percentage improvement it achieves over the average raw model error, and the best raw model error.

Variable	Best Scheme	Imp. Ave	Imp. Best
H_s	PWBC	36	14
Wind Speed	PWLC	31	18
Peak Period	PWBC	47	22

From the results for all variables, PWBC is generally performing better than PWLC as a compositing technique. despite individual models performing better with a learned linear correction than with a learned bias correction. Only in the case of wind speed, where linear corrected components show more twice the improvements of bias corrected components, do linear - corrected composites outperform bias - corrected composites.

For peak period for example, linear correction shows RMSE improvements of 27% over raw models, whilst bias correction shows only 18% improvement. Figure 3 shows the peak period bias as a function of peak period over three second bins. It can be seen from this figure that most individual model errors indicate a linear dependence on peak period, consistent with the positive response to this method of correction.

Despite this, PWBC proves a better compositing technique than PWLC (showing 47% and 42 % improvement respectively). A possible explanation for this lies in the fact that the compositing removes errors that are out of phase. Using bias - corrected components, opposing errors seen at high and low peak period for individual models cancel each other out, thus reducing the advantage gained from linear correction.

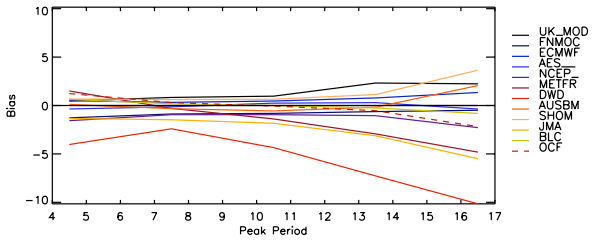

FIG. 3. Peak period bias as a function of period for individual models and PWBC

b. Varying Consensus Component Models

As discussed in section 3a the success of consensus forecasting techniques relies on the ability to remove systematic bias through learned bias or linear correction, as well as minimize random error through compositing. In the case of the latter, the effectiveness of this minimisation will bear a dependence on the number of component models making up the consensus. The intercomparison data set used in this work consists of ten models, however, in an operational setting, the number of models available in real time is likely far less than this. The following assesses the impact of the number of models on the performance of the consensus scheme.

Previous forecast experiments (e.g. Winkler et al. 1977) and theoretical studies (Clemen and Winkler 1985) studies have shown consensus forecasts usually improve rapidly from one to two components but the rate of

improvement drops asymptotically with further additions. Previous work using OCF techniques have had only a limited number of independent model components available for inclusion in consensus forecasts (3 for both Woodcock and Engel (2005) and Woodcock and Greenslade (2007)).

The large number of independent models available here provide an opportunity to address this question. A number of simulations were performed using subsets of the available models. Models were ranked according to their raw RMSE (see table 3). Consensus forecasts for the whole period were then constructed by consecutively dropping out models in increasing and decreasing order of quality.

Figure 4 shows the H_s RMSE of the PWBC as a function of the number of components models used to produce the consensus. For the "best n models" case, the worst models have been dropped out first, vice-versa for the "worst n models" case. The RMSE of the best and worst individual models have been included for reference.

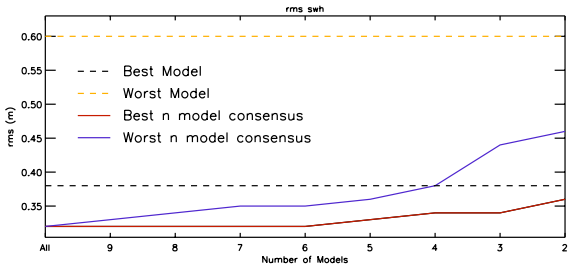

FIG. 4. 24 hour forecast H_s RMSE for PWBC as a function of the number of component models in the consensus. The best and worst n models are defined according the their RMSE. The RMSE of the best and worst individual models are included for reference.

For the "best models" case, increasing the number of models results in improvements in the consensus forecast only up to the inclusion of six models. Beyond this, the addition of more models, in this case the poorer performing models, does not add value to the forecast. However, a consensus forecast including the best model always does better than the best model on its own. For the case of the "worst models", a consensus of the two worst models does significantly better than the worst model raw forecast. Adding models to the consensus rapidly decreases error for the first five models, beyond which point, gains continue, though at a lesser rate. It is worth noting a consensus forecast using the worst four models is able to beat the best individual model.

Figure 5 shows the same plot for wind speed. As mentioned in section 4a, SHOM and ECMWF use the same winds, hence SHOM has been omitted from this analysis. For the most part, this plot shows a similar story to Figure 4. If anything, the benifits of increasing the number of consensus components beyond a minimum number is even less for wind speed than for H_s, with improvement in the "best models" case ceasing at five models. For the "worst models" case, the worst three model consensus beats the best individual model.

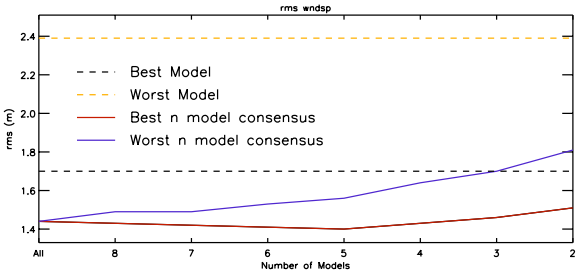

FIG. 5. As for figure 4 for wind speed.

This suggests that for the construction of a consensus forecast, ten models is unnecessary, and for practical purposes, little forecast skill is gained by adding further models beyond five or six. Although continuing to add better models does produce improvements, as seen in the "worst models" case in figures 4 and 5, in general, five or six seems optimal. This is of relevance in an operational environment where the cost of data retrieval and archiving must be weighed against potential gains in forecast ability.

When considering models to use for construction of a consensus forecast, not only must the quality of individual models be considered, but also model independence (Clemen and Winkler 1985). Compositing will most effectively remove component error in the case where individual components to have errors that are out of phase.

To illustrate this point, we consider here the simple case of constructing a consensus forecast from the ECMWF model and one other. We chose the model with the highest and lowest error correlations with the ECMWF model, namely SHOM (R = 0.85) and UKMO (R = 0.43) (error correlation coefficients between all raw models are given in Appendix D). Statistics for each of these raw models, as well as the PWBC consensus forecasts are given in table 11. From the raw model statistics, it is apparent that SHOM is a better performing model than the UKMO in this case. However, due to the low error correlation between ECMWF and UKMO, the consensus forecast using these models does better than that using the higher quality SHOM model. This is due to the fact that SHOM errors are highly correlated with ECMWF errors, reducing the impact of error minimisation due to compositing.

c. Extended Forecast Periods

Up until this point, only 24 hour forecasts have been considered. The following examines how these corrections perform for extended forecast periods out to five days. Of the ten models used in the previous section, only five produce forecasts out to five days, namely FNMOC,

Table 11. 48 hour H_s forecast statistics for ECMWF, UKMO and SHOM raw models, as well as PWBC results for different combinations of these models. R refers to the error correlation between the models used in the PWBC consensus.

Model		MAE	RMSE	XAE	SI	V%
ECMWF		0.31	0.45	2.28	0.18	87.77
UKMO		0.40	0.55	2.70	0.22	83.77
SHOM		0.34	0.47	2.59	0.19	86.81
		PWBC using specified components				
Components	R	MAE	RMSE	XAE	SI	V%
ECMWF and UKMO	0.43	0.29	0.41	2.25	0.17	89.66
ECMWF and SHOM	0.85	0.31	0.44	2.33	0.18	88.21

ECMWF, NCEP, SHOM and UKMO. In an operational environment, consensus forecasts would ideally be made with the maximum number of available models for each forecast period. However, it is the intention here to examine the relative performance of these schemes with increasing forecast period, and as such, a consistent model set across all forecast periods is desirable. To this end, the following results use only these five models for all forecast periods.

Figure 6 shows the RMSE growth with forecast period for each of these models as well as the BLC, PWBC and PWLC learned correction schemes for H_s, wind speed and peak period. Across all models, wind speed shows a rapid increase in forecast error from analysis to 24 hour forecast, steadily increasing throughout the remaining forecast period. For H_s, error increases only slightly between the analysis and the 24 hour forecast. This reflects the impact of data assimilation and the differing retention periods of the advantages gained. Assimilated information is retained longer in a wave model than an atmospheric model due to the longer temporal scales of variability. Also, the wave field consists of both wind sea and swell components. As the swell components are generated by winds earlier in the forecast period, a delay could be expected between when increasing errors in the winds are translated to corresponding errors in the wave forecast. Peak period error growth is slower than that of H_s and wind speed. This is likely due in part to the difficulties in accurately modelling peak period discussed in section 4a.

Learned correction schemes appear to retain their advantage throughout the forecast period, with results varying very little from 24 hour results.

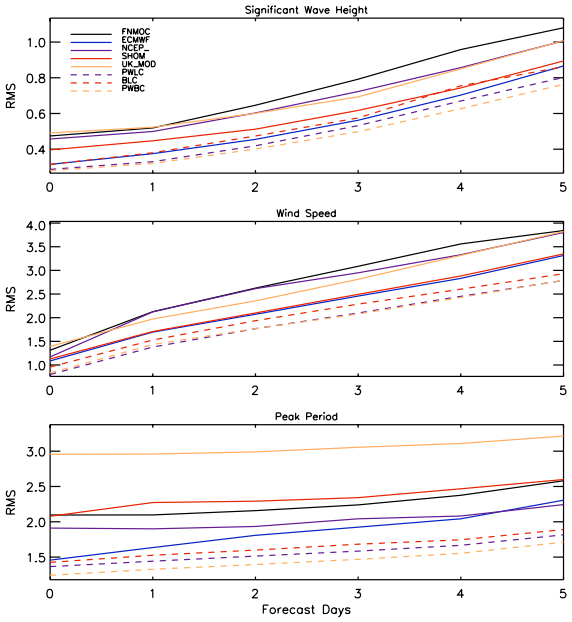

FIG. 6. RMSE growth with forecast period for model for which five day forecasts are available (FNMOC, ECMWF, NCEP, SHOM and UKMO) as well as PWBC, PWLC and BLC forecasts produced using these models.

5. Summary

The OCF scheme has been applied to forecasts of H_s, wind speed and peak period. Forecasts have been compiled using ten wave models at 14 buoy sites located around North America. A number of different correction techniques have been explored, including bias and linear correction of individual models, as well as composite forecasts constructed from equal weighted and performance weighted combinations of these bias and linearly corrected components.

Performance weighted composite schemes were found to be the best performers, with H_s and peak period achieving best results using bias corrected components, and wind speed using linearly corrected components. These composites resulted in improvements of 36%, 47% and 31% in RMSE values over the mean raw model components respectively. These improvements are in general, found to persist throughout the forecast period out to five days.

The large number of component models available has allowed the impact of the number of models used in the consensus forecast to be examined. It is found that little forecast skill is gained beyond five or six models. It is also noted that due to the nature of error minimisation during compositing, the degree of error correlation between models chosen for the composite must also be considered as well as the quality of the model with the aim being to maximise the degree to which component

errors are out of phase.

a. Further Work

In the work of Woodcock and Greenslade (2007), of the five models used for compositing, two were regional models nested within a third global model, resulting in highly correlated errors between there models. It was found that the best correction scheme was a so called composite of compostes, whereby these three models were first averaged, then OCF applied to this average and the remaining two models. One avenue that could be explored here would be an objective application of this idea, whereby models with highly correlated errors within the training period are first combined before inclusion in the consensus.

The potential also exist for OCF forecasts to be extended to grid based forecasts using altimeter observations.

Acknowledgements. The authors would like to thank Jean Bidlot at ECMWF for his continued efforts with the intercomparison project that provides the data for this work. We would also like to acknowledge all the institutions that contribute their data to this project.

APPENDIX A H_s **Statistics**

Table A1. 24 hour H_s statistics for all raw, bias corrected and linearly corrected models, as well as EWBC, PWBC, EWLC, PWLC and BLC forecasts.

MODEL	SCHEME	N	BIAS	MAE	RMSE	XAE	SI	V%
UKMO	RAW	4600	0.21	0.40	0.52	2.07	0.18	88.50
FNMOC	RAW	4600	0.09	0.39	0.52	2.29	0.19	88.36
ECMWF	RAW	4600	-0.01	0.27	0.38	1.70	0.15	92.18
MSC	RAW	4600	-0.09	0.32	0.44	2.31	0.17	89.85
NCEP	RAW	4600	0.16	0.38	0.51	2.07	0.18	88.94
METFR	RAW	4600	0.23	0.46	0.60	2.45	0.22	83.66
DWD	RAW	4600	-0.07	0.38	0.51	2.22	0.19	87.41
AUSBM	RAW	4600	-0.11	0.42	0.57	2.53	0.23	82.89
SHOM	RAW	4600	-0.01	0.33	0.44	2.09	0.17	90.19
JMA	RAW	4600	-0.15	0.44	0.59	2.50	0.23	82.21
UKMO	BC	4600	0.01	0.34	0.46	1.92	0.18	88.65
FNMOC	BC	4600	-0.01	0.33	0.46	2.22	0.19	88.37
ECMWF	BC	4600	-0.01	0.26	0.36	1.69	0.15	91.82
MSC	BC	4600	0.01	0.30	0.41	2.20	0.17	89.77
NCEP	BC	4600	0.00	0.32	0.44	2.00	0.18	88.58
METFR	BC	4600	0.01	0.39	0.54	2.39	0.22	83.43
DWD	BC	4600	-0.01	0.34	0.45	1.98	0.18	87.87
AUSBM	BC	4600	-0.01	0.39	0.53	2.50	0.22	83.16
SHOM	BC	4600	0.01	0.30	0.41	2.12	0.17	90.24
JMA	BC	4600	-0.02	0.40	0.54	2.40	0.22	82.28
UKMO	LC	4600	-0.00	0.33	0.44	2.15	0.18	88.20
FNMOC	LC	4600	0.01	0.32	0.45	2.15	0.18	87.90
ECMWF	LC	4600	-0.00	0.26	0.36	1.74	0.15	91.53
MSC	LC	4600	0.00	0.30	0.42	2.05	0.17	89.24
NCEP	LC	4600	0.00	0.31	0.43	2.11	0.18	88.04
METFR	LC	4600	0.00	0.39	0.54	2.53	0.22	82.42
DWD	LC	4600	-0.00	0.34	0.47	2.48	0.19	86.86
AUSBM	LC	4600	0.01	0.39	0.54	2.79	0.22	81.99
SHOM	LC	4600	0.00	0.30	0.41	2.22	0.17	89.55
JMA	LC	4600	-0.01	0.40	0.55	2.48	0.23	81.72
BLC	na	4600	0.00	0.27	0.38	1.96	0.16	91.08
EWBC	Composite	4600	-0.00	0.25	0.34	1.52	0.14	93.14
PWBC	Composite	4600	-0.00	0.24	0.32	1.46	0.14	93.57
EWLC	Composite	4600	0.00	0.26	0.35	1.67	0.15	92.32
PWLC	Composite	4600	0.00	0.25	0.34	1.56	0.14	92.89
PERSIST	na	4600	0.00	0.87	1.23	5.74	0.51	59.77

APPENDIX B **Wind Speed Statistics**

Table B1. 24 hour wind speed statistics for all raw, bias corrected and linearly corrected models, as well as EWBC, PWBC, EWLC, PWLC and BLC forecasts.

MODEL	SCHEME	N	BIAS	MAE	RMSE	XAE	SI	V%
UKMO	RAW	4291	0.81	1.52	1.98	9.42	0.23	80.36
FNMOC	RAW	4291	0.55	1.64	2.14	9.79	0.26	73.12
ECMWF	RAW	4291	0.46	1.29	1.71	8.67	0.21	81.24
MSC	RAW	4291	0.49	1.52	2.02	8.68	0.25	74.41
NCEP	RAW	4291	0.71	1.65	2.15	8.62	0.26	74.52
METFR	RAW	4291	0.16	1.44	1.91	9.16	0.24	75.90
DWD	RAW	4291	0.18	1.46	1.92	8.56	0.24	73.91
AUSBM	RAW	4291	0.71	1.86	2.40	10.30	0.29	66.78
SHOM	RAW	4291	0.38	1.28	1.69	8.73	0.21	81.04
JMA	RAW	4291	0.68	1.63	2.12	9.31	0.25	75.79
UKMO	BC	4291	0.00	1.31	1.78	9.29	0.23	79.79
FNMOC	BC	4291	-0.01	1.54	2.04	9.46	0.27	72.22
ECMWF	BC	4291	-0.01	1.19	1.61	8.66	0.21	80.97
MSC	BC	4291	0.00	1.42	1.91	9.07	0.25	74.25
NCEP	BC	4291	0.02	1.47	1.96	8.47	0.26	74.48
METFR	BC	4291	-0.04	1.40	1.86	8.85	0.24	75.35
DWD	BC	4291	0.01	1.41	1.88	8.63	0.25	74.02
AUSBM	BC	4291	-0.07	1.70	2.26	10.69	0.30	65.35
SHOM	BC	4291	-0.01	1.19	1.61	8.84	0.21	80.82
JMA	BC	4291	-0.00	1.47	1.96	9.33	0.26	75.84
UKMO	LC	4291	0.01	1.15	1.53	7.88	0.20	79.52
FNMOC	LC	4291	0.01	1.38	1.80	8.31	0.24	71.72
ECMWF	LC	4291	0.01	1.12	1.50	7.63	0.20	80.28
MSC	LC	4291	0.02	1.30	1.71	8.03	0.22	73.97
NCEP	LC	4291	0.01	1.32	1.72	7.09	0.23	74.16
METFR	LC	4291	-0.00	1.29	1.70	7.87	0.22	74.79
DWD	LC	4291	0.02	1.32	1.73	7.57	0.23	73.87
AUSBM	LC	4291	0.01	1.53	2.01	9.20	0.26	64.86
SHOM	LC	4291	0.02	1.13	1.51	7.80	0.20	80.02
JMA	LC	4291	0.01	1.27	1.68	7.95	0.22	75.69
BLC	na	4291	0.04	1.14	1.53	7.66	0.20	79.79
PWBC	Composite	4291	-0.01	1.08	1.43	6.71	0.19	84.26
EWBC	Composite	4291	-0.02	1.10	1.45	6.59	0.19	83.82
EWLC	Composite	4291	0.01	1.08	1.42	6.23	0.19	82.81
PWLC	Composite	4291	0.01	1.05	1.39	6.28	0.18	83.42
PERSIST	na	4291	0.01	3.16	4.04	13.14	0.53	29.68

APPENDIX C **Peak Period Statistics**

Table C1. 24 hour peak period statistics for all raw, bias corrected and linearly corrected models, as well as EWBC, PWBC, EWLC, PWLC and BLC forecasts.

MODEL	SCHEME	N	BIAS	MAE	RMSE	XAE	SI	V%
UKMO	RAW	4259	0.98	2.10	2.91	10.74	0.28	17.57
FNMOC	RAW	4259	-0.79	1.47	2.10	10.72	0.20	41.38
ECMWF	RAW	4259	0.34	0.97	1.64	10.21	0.17	56.66
MSC	RAW	4259	-0.01	1.05	1.73	9.66	0.19	48.53
NCEP	RAW	4259	-0.92	1.31	1.90	9.21	0.18	49.07
METFR	RAW	4259	-0.72	1.46	1.94	7.57	0.19	38.50
DWD	RAW	4259	-3.66	3.81	4.64	12.69	0.30	4.12
AUSBM	RAW	4259	-0.21	1.79	2.68	13.15	0.28	17.45
SHOM	RAW	4259	0.61	1.31	2.26	11.97	0.23	35.48
JMA	RAW	4259	-1.54	2.00	2.58	9.51	0.22	23.78
UKMO	BC	4259	0.02	1.84	2.51	10.10	0.27	23.16
FNMOC	BC	4259	0.03	1.14	1.84	10.18	0.20	43.17
ECMWF	BC	4259	0.15	0.97	1.60	10.01	0.17	56.39
MSC	BC	4259	-0.01	1.03	1.67	9.49	0.18	48.85
NCEP	BC	4259	-0.10	1.03	1.66	8.79	0.18	49.17
METFR	BC	4259	-0.05	1.30	1.74	7.42	0.19	39.38
DWD	BC	4259	-0.05	1.93	2.50	8.91	0.27	8.32
AUSBM	BC	4259	0.21	1.64	2.56	13.53	0.28	21.49
SHOM	BC	4259	0.30	1.30	2.13	11.59	0.23	37.53
JMA	BC	4259	-0.08	1.27	1.83	8.56	0.20	34.25
UKMO	LC	4259	0.01	1.45	1.96	7.96	0.21	25.54
FNMOC	LC	4259	0.03	1.14	1.69	8.65	0.18	43.10
ECMWF	LC	4259	0.02	0.99	1.50	8.87	0.16	54.74
MSC	LC	4259	-0.04	1.10	1.63	8.53	0.18	45.17
NCEP	LC	4259	0.00	1.07	1.60	8.33	0.18	47.00
METFR	LC	4259	-0.00	1.32	1.75	7.16	0.19	38.72
DWD	LC	4259	-0.02	1.56	2.04	7.71	0.22	17.34
AUSBM	LC	4259	0.07	1.40	2.01	10.42	0.22	23.58
SHOM	LC	4259	0.03	1.22	1.78	9.74	0.19	38.07
JMA	LC	4259	0.07	1.29	1.80	8.45	0.20	35.17
BLC	na	4259	0.04	0.98	1.53	9.30	0.17	53.62
PWBC	Composite	4259	0.04	0.87	1.29	6.99	0.14	66.74
EWBC	Composite	4259	0.06	0.96	1.37	6.46	0.15	63.65
EWLC	Composite	4259	0.02	1.09	1.51	7.04	0.16	57.20
PWLC	Composite	4259	0.01	1.00	1.42	7.26	0.15	60.85
PERSIST	na	4259	0.00	1.87	2.50	9.58	0.27	35.42

APPENDIX D **Raw model error correlations**

Table D1. Error correlations between each model

	UKMO	FNMOC	ECMWF	MSC	NCEP	METFR	DWD	AUSBM	SHOM	JMA
UKMO	1.00	0.41	0.43	0.40	0.38	0.51	0.51	0.40	0.49	0.23
FNMOC		1.00	0.54	0.42	0.78	0.22	0.46	0.47	0.59	0.21
ECMWF			1.00	0.66	0.56	0.57	0.62	0.57	0.85	0.45
MSC				1.00	0.43	0.54	0.56	0.49	0.61	0.37
NCEP					1.00	0.24	0.41	0.42	0.58	0.23
METFR						1.00	0.58	0.57	0.48	0.45
DWD							1.00	0.75	0.66	0.43
AUSBM								1.00	0.57	0.46
SHOM									1.00	0.30
JMA										1.00

REFERENCES

Bidlot, J., and M. Holt, 2006: Verification of operational global and regional wave forecasting systems against measurements from moored buoys. Technical Report 30. WMO/TDNo.1333., JCOMM.

Bidlot, J. R., D. J. Holmes, P. A. Wittmann, R. Lalbeharry, and H. S. Chen, 2002: Intercomparison of the performance of operational ocean wave forecasting systems with buoy data. *WEATHER AND FORECASTING*, **17**, 287–310.

Bidlot, J. R., J.-G. Li, P. A. Wittmann, M. Fauchon, H. S. Chen, J.-M. Lefevre, T. Bruns, D. J. M. Greenslade, F. Ardhuin, N. Kohno, S. Park, and M. Gomez, 2007: Inter-comparison of operational wave forecasting systems. *Proceedings of the 8th International Workshop on wave hindcasting and forecasting, Oahu, Hawaii, USA, November 2007.*.

Booij, N., R. Ris, and L. Holthuijsen, 1999: A third-generation wave model for coastal regions, Part I, Model description and validation. *Journal of Geophysical Research*, **104**, 7649–7666.

Caires, S., and A. Sterl, 2003: Validation of ocean wind and wave data using triple collocation. *Journal of Geophysical Research*, **108**, 3098.

Caires, S., A. Sterl, J. R. Bidlot, N. Graham, and V. Swail, 2004: Intercomparison of different wind-wave reanalyses. *Journal of Climate*, **17**, 1893–1913.

Clemen, R., and R. Winkler, 1985: Limits for the Precision and Value of Information from Dependent Sources. *Operations Research*, **33**, 427–442.

Durrant, T., D. Greenslade, and I. Simmonds, Submitted: Validation of jason-1 and envisat remotely sensed wave heights. *J. Atmos. Oc. Tech*.

Faugere, Y., J. Dorandeu, F. Lefevre, N. Picot, and P. Femenias, 2006: Envisat ocean altimetry performance assessment and cross-calibration. *Sensors*, **6**, 100–130.

Glahn, H., and D. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11**, 1203–1211.

Janssen, P. A. E. M., B. Hansen, and J. R. Bidlot, 1997: Verification of the ECMWF wave forecasting system against buoy and altimeter data. *Weather and Forecasting*, **12**, 763–784.

Komen, G., 1994: *Dynamics and Modelling of Ocean Waves*. Cambridge University Press.

Queffeulou, P., 2004: Long-Term Validation of Wave Height Measurements from Altimeters. *Marine Geodesy*, **27**, 495–510.

Tolman, H., 1991: A Third-Generation Model for Wind Waves on Slowly Varying, Unsteady, and Inhomogeneous Depths and Currents. *Journal of Physical Oceanography*, **21**, 782–797.

Tolman, H. L., 2002: Validation of WAVEWATCH III version 1.15 for a global domain. Technical Note Nr. 213 p 33, NCEP.

WAMDI-Group, 1988: The WAM model - A third generation ocean wave prediction model. *J. Phys. Oceanogr*, **18**, 1775–1810.

Winkler, R., A. Murphy, and R. Katz, 1977: The consensus of subjective probability forecasts: Are two, three,... heads better than one? Preprints. *Fifth Conf. on Probability and Statistics*, 57–62.

Wonnacott, T., and R. Wonnacott, 1972: Introductory Statistics p. 287.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Weather and forecasting*, **20**, 101–111.

Woodcock, F., and D. J. M. Greenslade, 2007: Consensus of numerical model forecasts of significant wave heights. *Weather and Forecasting*, **22**, 792–803.