Consensus Forecasts of Modeled Wave Parameters

TOM H. DURRANT, FRANK WOODCOCK, AND DIANA J. M. GREENSLADE

Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Victoria, Australia

(Manuscript received 25 March 2008, in final form 22 August 2008)

ABSTRACT

The use of numerical guidance has become integral to the process of modern weather forecasting. Using various techniques, postprocessing of numerical model output has been shown to mitigate some of the deficiencies of these models, producing more accurate forecasts. 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 (T_p), and 10-m wind speed (U_{10}) from 10 numerical wave models, at 14 buoy sites located around North America. Results show the best forecast is achieved with a weighted average of bias-corrected components for both H_s and T_p, while a weighted average of linear-corrected components gives the best results for U_{10}. For 24-h forecasts, improvements of 36%, 47%, and 31%, in root-mean-square-error values over the mean raw model components are achieved, or 14%, 22%, and 18% over the best individual model. Similar gains in forecast skill are retained out to 5 days. By reducing the number of models used in the construction of consensus forecasts, it is found that little forecast skill is gained beyond five or six model components, with the independence of these components, as well as individual component's quality, being important considerations. It is noted that for H_s it is possible to beat the best individual model with a composite forecast of the worst four.

1. Introduction

Increasing computational power, improved modeling techniques, and increased availability of observations have facilitated rapid development of numerical weather prediction (NWP) capabilities in recent years. Deficiencies remain, however, due to factors such as imperfect model physics and uncertainties in initial and boundary conditions (Cheng and Steenburgh 2007). In addition, grid-point values of model output fields represent an area dependent upon the spatial resolution of the model grid. Interpolation of this output to specific locations may result in systematic biases due to unresolved local effects (Engel and Ebert 2007).

Postprocessing techniques aim to reduce these systematic biases. The widely used model output statistics (MOS), for example, uses multiple linear regression based on model output and previous observations to provide improved forecasts at specific locations (Glahn and Lowry 1972). A major drawback to MOS is the long training dataset required. This results in a poor ability to adapt to new sites, additional models, and upgrades to existing models. The increasing frequency of numerical and observational system changes suggests the importance of direct model output (DMO) will increase relative to MOS (Woodcock and Engel 2005).

When guidance is available from a number of different models, consensus techniques, which combine available guidance, have been found to provide improvements over individual DMOs (Hibon and Evgeniou 2005). The operational consensus forecast (OCF) scheme of Woodcock and Engel (2005) is one example of how a consensus forecast can be constructed. It 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.

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

DOI: 10.1175/2008WAF2222143.1

In operational daily weather prediction at the Australian Bureau of Meteorology (the Bureau), OCF combines both operationally available MOS forecasts and bilinearly interpolated DMO 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.

It is worth briefly comparing and contrasting OCF to ensemble predictions, such as those produced at the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hoffschildt et al. 2000; Janssen 2000) and the National Centers for Environmental Prediction (NCEP) (Chen 2006). For an ensemble wave prediction, divergent forecasts can be produced by perturbing the initial conditions or wind forcing, for example. These divergent solutions can then be combined to produce greater reliability in the solution, as well as providing probabilistic information about the spread of conceivable outcomes. OCF provides a single, deterministic forecast only. The strength of OCF comes from its ability to remove systematic bias in the model, something that is retained in an ensemble forecast. Further discussion on the application of ensemble forecasting to wave models can be found in Farina (2002).

Woodcock and Greenslade (2007) investigated the application of OCF techniques to wave forecasts. They employed OCF to generate 24-h predictions of significant wave height (H_s) at 18 observation locations around Australia. Among the broad conclusions reached 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, 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 here with 10 independent models from the major forecasting centers used for compositing. 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 the peak period (T_p) from these same models, as well as the 10-m wind speed (U_{10}) used to force them, is investigated at these same sites. The analysis is also extended to cover increased forecast periods out to 5 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.

The model and observational data are examined in section 2, general descriptions of the OCF techniques and of the specific application used here are described in sections 3 and 4, results are presented in section 5, and finally section 6 contains a summary of the work.

2. Data

For the past 6 yr, a monthly exchange of ocean wave model data has been taking place between the major forecasting centers around the world (Bidlot et al. 2002). What started as a cooperation between the ECMWF, the Met Office (UKMO), the Fleet Numerical Meteorology and Oceanography Center (FNMOC), the Meteorological Service of Canada (MSC), and NCEP has now grown to include, in chronological order of participation, Deutscher Wetherdienst (DWD), the Bureau, Météo-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 center 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 U_{10} are available. These data are then analyzed at ECMWF, with various summary plots and statistics produced. These processed products, as well as the collated raw data for all centers are then made available to all participants. It is this dataset that provides the basis for this work.

a. Observational data

Observational data used here 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 (e.g., Janssen et al. 1997; Caires and Sterl 2003; Caires et al. 2004) and altimeter (e.g., Tolman 2002; Queffeulou 2004; Faugere et al. 2006) data. As part of the collocation process performed at model C, these data undergo a quality control process to remove suspect observations. Wind speeds are adjusted to 10-m height, and spatial and temporal scales are made comparable by averaging the hourly observations in time windows of 4 h, 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

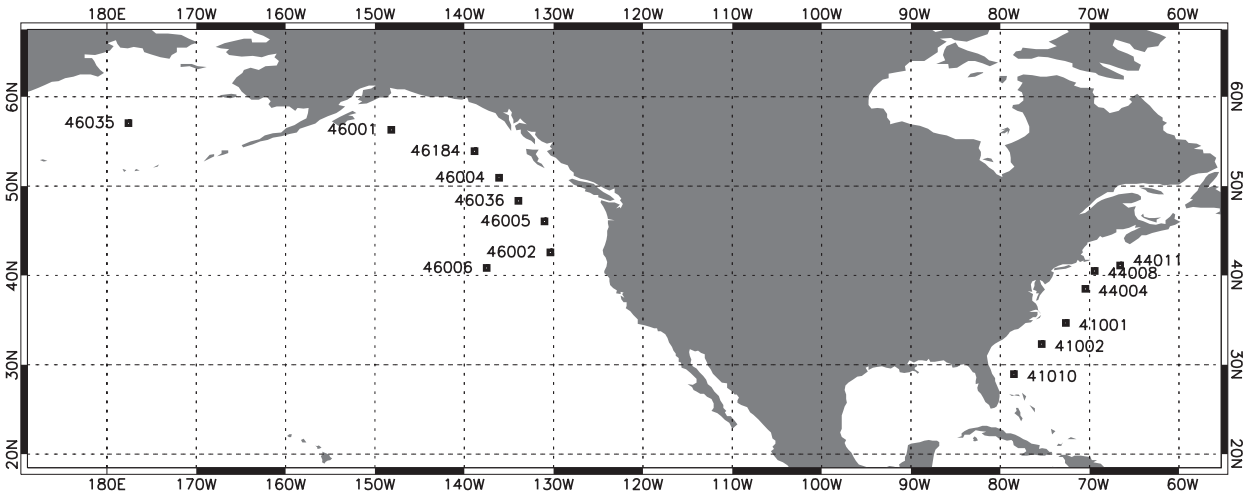

FIG. 1. Location of buoys used in this study (see Table 1).

contain short time series of historical data. Others have only a subset of model data available at the site. Hence, in order to achieve a clean dataset with the maximum number of models, a subset of 14 buoys was chosen here for which all participating models are present. KMA and PRTOS joined the intercomparison only recently (July 2007) and are 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 Fig. 1, with details 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 25 m (Booij et al. 1999).

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. 2009) suggests that systematic differences may exist between these two networks, with MEDS-reported H_s values being 10% lower 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

The provision of the full details of all these models is impractical here, and further references can be found in Bidlot et al. (2007). Two models are dominant, namely the third-generation Wave Model (WAM; WAMDI Group 1988; Komen 1994) and WAVEWATCH III (WW3; Tolman 1991). Operational versions of these models at each forecasting center have, however, undergone many independent changes and tunings. All models also have different wind forcings, 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. Forecasts for all models are produced 12 hourly.

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. To this end, models are referred to by generic names: model A, model B, etc., and comments regarding the performance of individual models are made only where it is relevant within the context of a composite scheme's performance. Comparative analysis of model performance based on this dataset can be found in Bidlot et al. (2002) and more recently Bidlot et al. (2007).

TABLE 1. Details of the buoys used in this study. WMO refers to the number assigned by the World Meteorological Organization.

WMO	Name	Owner	Lat (°N)	Lon (°W)	Depth (m)
41001	East Hatteras	NDBC	34.68	72.66	4427
41002	South 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

3. OCF methodology

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 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 minimized 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:

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

where Q_1, Q_2, and Q_3 are the error sample's first, second, and third quartiles, respectively. BES is robust with respect to extreme values but represents the bulk of the common results. Normalized 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

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

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

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

Breaking the forecasts (f_i) into the aforementioned components,

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

Gathering terms, this becomes

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

The final two terms in Eq. (5) highlight the importance of the bias removal and weighting schemes. Characterization of the random nature of the error distributions, as part of the weighting scheme, aids minimization of the random errors via compositing with highly variable models penalized for their reduced reliability.

4. Experimental setup

A number of corrected forecasting techniques are explored based around the OCF technique described in the previous section. The first class of corrections used here is made up of those in which the model forecast is corrected according to a training set based on that particular model. These types of corrections do not use any compositing. Two such methods are investigated: a simple bias correction, and a least squares linear-regression correction. In the case of the former, a bias correction is calculated from the training period using BES and applied to the next forecast. For the latter, the bias correction is replaced by a linear-regression equation.

The second class of corrections involves compositing. Several forms of compositing are 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 are 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, forecasts are generated 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-corrected, linear-regression-corrected, and BLC comparisons are undertaken over corresponding, matching events (i.e., identical verifying sets and training windows).

The effects of varying the training period were investigated by Woodcock and Greenslade (2007) by increasing the training window in steps of four from 1 to 59 forecast events. They found that bias-corrected methods stabilized by 9 events, and linearly corrected methods by 13 events. For the bulk of their work, a fixed training window of 29 events, or 14.5 days, was used. To maintain consistency, the same is used here.

TABLE 2. Statistics for 24-h H_s forecasts for all raw models used in this study. Statistics are based on 4600 individual observations. The best performing model for each statistic is indicated in boldface.

	RMSE							
Model	Bias	MAE (m)	Median (m)	5%	95%	XAE (m)	SI	%V
A	0.21	0.40	0.54	0.53	0.56	2.07	0.18	88.50
B	0.09	0.39	0.53	0.52	0.55	2.29	0.19	88.36
C	**−0.01**	**0.27**	**0.38**	0.37	0.39	**1.70**	**0.15**	**92.18**
D	−0.09	0.32	0.45	0.44	0.47	2.31	0.17	89.85
E	0.16	0.38	0.52	0.51	0.53	2.07	0.18	88.94
F	0.23	0.46	0.61	0.60	0.63	2.45	0.22	83.66
G	−0.07	0.38	0.51	0.50	0.52	2.22	0.19	87.41
H	−0.11	0.42	0.57	0.56	0.59	2.53	0.23	82.89
I	**−0.01**	0.33	0.45	0.44	0.46	2.09	0.17	90.19
J	−0.15	0.44	0.60	0.59	0.61	2.50	0.23	82.21

5. Results

We begin by examining the results for 24-h H_s forecasts. The same correction techniques are then extended to 24-h forecasts of U_{10} and T_p. 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.

While these studies consider results in various regions, all buoys are considered together here. For the purpose of intercomparison and model diagnostics, the separation provides further insight into sources of error by comparing wind sea versus swell-dominated areas, for example, or where various sheltering and subgrid-scale processes are of differing importance. For this work, while the performance of the statistical scheme will differ with the quality of the input models, due to its nonphysical nature, little is gained by examining regions separately. The same is also true for examining seasonal variations in error.

Verification statistics include bias, MAE, root-mean-square error (RMSE), maximum absolute error (XAE), the scatter index (SI), and the percentage of explained variance (%V = 100 × the square of the correlation between forecast and observation) defined as follows:

$$\text{bias} = \frac{1}{N}\sum_{j=1}^{N} F_j - O_j, \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(F_j - O_j)^2}, \tag{7}$$

$$\text{SI} = \frac{\sqrt{\frac{1}{N}\sum_{j=1}^{N}[(F_j - \overline{F}) - (O_j - \overline{O})]^2}}{\overline{O}}, \quad \text{and} \tag{8}$$

TABLE 3. Percentage improvements for 24-h H_s forecasts for various correction schemes over the average 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). The best performing correction scheme for each statistic is indicated in boldface.

Model	MAE (m)	RMSE (m)	XAE (m)	SI	%V
Avg raw	0.38	0.51	2.22	0.19	87.42
	Improvement over average of raw models (%)				
Avg BC	11	9	4	1	−0
Avg 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

$$\%\text{V} = 100 \times \left[\frac{\sum_{j=1}^{N}(F_j - \overline{F})(O_j - \overline{O})}{\sqrt{\sum_{j=1}^{N}(F_j - \overline{F})^2 (O_j - \overline{O})^2}}\right]^2, \tag{9}$$

where F_j is the forecast value, O_j is the observed value, N is the number of independent forecasts, and an overbar represents the mean value. For RMSE, the median and its 90% confidence interval are generated using the bootstrap method (Efron and Tibshirani 1991). This involves repeated reestimation using random samples with replacement from the original data, with 10^6 iterations performed here.

a. 24-h forecast results

Verification statistics for 24-h forecasts for each model are presented in Table 2. Model C 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 14 buoys used here.

Table 3 shows the improvements of various correction schemes over the average raw model error. 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, model C remains the best model after bias and linear correction. Due to its negligible overall bias, little improvement is made on this model by applying these corrections, resulting in little difference between the numbers seen here. Of the composite schemes, those that are bias

TABLE 4. As in Table 2 but for U_{10}. Statistics are based on 4830 individual forecasts.

	RMSE							
Model	Bias	MAE (m)	Median (m)	5%	95%	XAE (m)	SI	%V
A	0.81	1.52	2.01	1.96	2.07	9.42	0.23	80.36
B	0.55	1.64	2.16	2.10	2.21	9.79	0.26	73.12
C	0.46	1.29	1.73	1.68	1.78	8.67	**0.21**	**81.24**
D	0.49	1.52	2.04	1.98	2.09	8.68	0.25	74.41
E	0.71	1.65	2.18	2.13	2.24	**8.62**	0.26	74.52
F	**0.16**	1.44	1.91	1.86	1.97	9.16	0.24	75.90
G	0.18	1.46	1.93	1.89	1.98	8.56	0.24	73.91
H	0.71	1.86	2.42	2.37	2.48	10.30	0.29	66.78
I	0.38	**1.28**	**1.71**	1.66	1.76	8.73	**0.21**	81.04
J	0.68	1.63	2.15	2.10	2.20	9.31	0.25	75.79

corrected are outperforming linear correction 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), who 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 the limited number of models included in the composite, and a lack of independence among these models.

Similar to Table 2 for H_s, Table 4 shows raw model statistics for U_{10}. It should be noted that models I and C use the same winds, though at different resolutions (0.5° spatial and 6-hourly temporal resolutions for model I; 40 km and 15-min resolutions for model C; F. Ardhuin 2006, personal communication). For this reason, the statistics for these winds are very similar, though perhaps surprisingly, the MAE and RMSE for model I winds are in fact lower for these buoys than those for model C. This, it seems, is due to an increased positive bias in the model C winds, with SI being the same for both centers. Bias-corrected results yield little difference between the two in terms of MAE and RMSE, while model C 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. Its presence suggests that a learned bias correction may have a positive impact on the U_{10} forecast.

Percentage improvements of the various correction schemes relative to the average raw model error are shown in Table 5. Unlike H_s, which showed similar improvements for bias and linear correction schemes, U_{10} performs better under a linear correction with more than double the improvement in the average linear-corrected model RMSE (16%) compared to the improvement of the average bias-corrected model RMSE (6%). It is unclear why this is the case. For U_{10}, the better response to a linear correction is consistent with the models not adequately capturing the high and low extremes. One possibility is that the damped response in the wave model may result in this tendency in the forcing winds being reduced somewhat, making a simple bias correction perform relatively better for the former. 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.

TABLE 5. As in Table 3 but for U_{10}.

Model	MAE (m)	RMSE (m)	XAE (m)	SI	%V
Avg raw	1.53	2.00	9.12	0.25	75.71
	Improvement over average of raw models (%)				
Avg BC	8	6	−0	−1	−1
Avg 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

Raw model statistics for T_p are shown in Table 6. Once again, model C is the best model here, with the lowest MAE, RMSE, and SI, and the highest %V values. The %V values are typically far lower than those seen for H_s and U_{10}. This reflects the difficulties associated with the verification of this variable. Peak period

TABLE 6. As in Table 2 but for T_p. Statistics are based on 4259 individual forecasts.

	RMSE							
Model	Bias	MAE (m)	Median (m)	5%	95%	XAE (m)	SI	%V
A	0.98	2.10	3.08	3.00	3.16	10.74	0.28	17.57
B	−0.79	1.47	2.11	2.03	2.18	10.72	0.20	41.38
C	0.34	**0.97**	**1.68**	1.60	1.76	10.21	**0.17**	**56.66**
D	**−0.01**	1.05	1.73	1.66	1.81	9.66	0.19	48.53
E	−0.92	1.31	1.93	1.86	1.99	9.21	0.18	49.07
F	−0.72	1.46	1.99	1.94	2.04	**7.57**	0.19	38.50
G	−3.66	3.81	4.76	4.68	4.84	12.69	0.30	4.12
H	−0.21	1.79	2.75	2.65	2.84	13.15	0.28	17.45
I	0.61	1.31	2.36	2.26	2.47	11.97	0.23	35.48
J	−1.54	2.00	2.61	2.55	2.67	9.51	0.22	23.78

TABLE 7. As in Table 3 but for T_p.

Model	MAE (m)	RMSE (m)	XAE (m)	SI	%V
Avg raw	1.73	2.44	10.54	0.22	33.25
Improvement over average of raw models (%)					
Avg BC	22	18	6	4	9
Avg 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

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 T_p values. For example, in the case of a bimodal spectrum with two near-equal peaks, small errors in the energy associated with either part of the spectrum can lead to a large jump in the T_p as it moves from one peak to the other.

Percentage improvements coming from the same correction schemes for T_p are shown in Table 7. Individual models seem to respond well to learned correction schemes, with bias and linear corrections resulting in average improvements of 18% and 27%, respectively, in RMSE over the average raw model error. 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, though this is likely due to the high quality of the best model compared to the average raw model, which would be expected to be featured heavily in the BLC forecast.

In the case of each variable (H_s, U_{10}, and T_p), results have been presented here in terms of their improvement over the average raw model error. 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 T_p for example, PWBC achieves an impressive 47% improvement over the average raw model 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, in order to gauge the true gains being achieved through compositing, it is important to consider not only the improvement over the average component models, but also those relative to the best individual component model. To separate the benefits of compositing verses those from bias correction alone, it is also worth considering improvement over the best corrected component model.

TABLE 8. Shown are the best performing correction schemes for each variable, and the RMSE percentage improvement each achieves over the average raw model error, the best raw model error, and the best corrected model (bias correction for H_s and T_p and linear correction for U_{10}).

Variable	Best scheme	Improvement over avg error (%)	Improvement over best raw model error (%)	Improvement over best corrected model (%)
H_s	PWBC	36	14	11
U_{10}	PWLC	31	18	7
T_p	PWBC	47	22	19

Table 8 shows, for each variable, the best performing correction scheme, the improvement over the average error of the raw components, the improvement over the best raw model, and the improvement over the best corrected model with respect to RMSE. The best corrected model is a bias-corrected component in the cases of H_s and T_p and a linear-corrected component in the case of U_{10}. Relative to the average error, H_s and T_p see far greater gains than those relative to the best model, while this is less so for U_{10}. This is a reflection of the significant gains obtained from individual component correction for U_{10} seen in Table 5. For this reason, gains over the best corrected model are far less than those for the best raw model. These results reflect the relative spread in the quality of the modeled variable, with more consistency seen in the modeled wind fields across the various institutions than the wave model variables.

b. Number of component models

As discussed in section 3 the success of consensus forecasting techniques relies on their ability to remove systematic bias through learned bias or linear correction, as well as the ability to minimize random error through compositing. In the case of the latter, the effectiveness of this minimization will depend on the number of component models making up the consensus. The intercomparison dataset used in this work consists of 10 models. 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) have shown that 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 has only had a limited number of independent model components

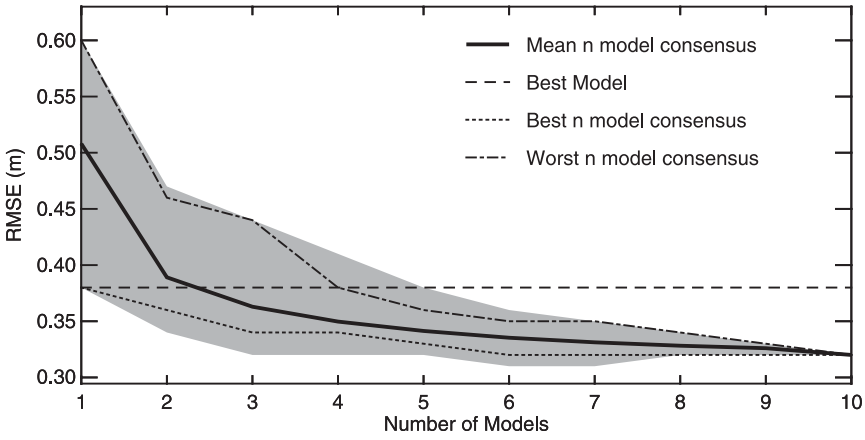

FIG. 2. The 24-h forecast H_s RMSE for PWBC forecasts as a function of the number of component models in the consensus. The best and worst n models are defined according to their RMSEs. The shading shows the spread of all possible combinations of n models from the 10 available models.

from different centers available for inclusion in consensus forecasts [three for both Woodcock and Engel (2005) and Woodcock and Greenslade (2007)].

The large number of independent models available here provides an opportunity to explore this issue. A number of simulations were performed using subsets of the available models. Figure 2 shows the RMSE of PWBC forecasts as a function of the number of models included in the consensus. For each step n, all possible combinations of n models from the available 10 models were used to produce a set of composite forecasts. For example, in the case of 4 models, there are 210 unique combinations of the 10 available models. The gray area shows the spread of this set of composite forecasts, with the mean error indicated. In addition, models were ranked according to their raw RMSE (see Table 2), and consensus forecasts for the whole period were then constructed by consecutively dropping out models in increasing and decreasing order of quality. These are indicated as the *best n models* and *worst n models*.

As in Winkler et al. (1977) and Clemen and Winkler (1985), significant improvements are seen from one to two components, with gains asymptoting beyond this. Note that the best models and worst models lines do not correspond to the extremes of the spread of all possible combinations. This is due to the importance of uncorrelated errors in the production of a good consensus forecast, which is discussed further below. 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. A consensus forecast including the best model always does better than the best model on its own. For the case of the worst models, adding models to the consensus rapidly decreases the error for the first five models, beyond which point gains continue, though at a lesser rate. It is worth noting that a consensus forecast using the worst four models is able to beat the best individual model. Though this relationship will vary somewhat for any given forecast parameter depending on the spread of the ensemble, this is a very encouraging result.

Figure 3 shows the same plot for U_{10}. As mentioned in section 5a, models I and C use the same winds; hence, model I has been omitted from this analysis. PWLC is shown here, as this produces the best forecast in the case of U_{10}. This plot shows many similarities with Fig. 2. The gain in forecast skill from one model to two is greater however, with gains beyond this dropping off faster. The benefits of increasing the number of consensus components beyond a minimum number is even less for U_{10} than for H_s, with improvement in the best models case ceasing at five models. For the worst models case, the worst two-model consensus beats the best individual model.

This suggests that for the construction of a consensus forecast, 10 models is unnecessary, and for practical purposes, little forecast skill is gained by adding further models beyond 5 or 6. Although continuing to add better models does produce improvements, as seen in the worst models case in Figs. 2 and 3, the *mean n model consensus* from these figures suggests five or six is optimal. This is of relevance in an operational environment where the cost of data retrieval and archiving must be weighed against any potential gains in forecast ability. From the spread of all possible combinations, it seems that, in certain circumstances, adding more models

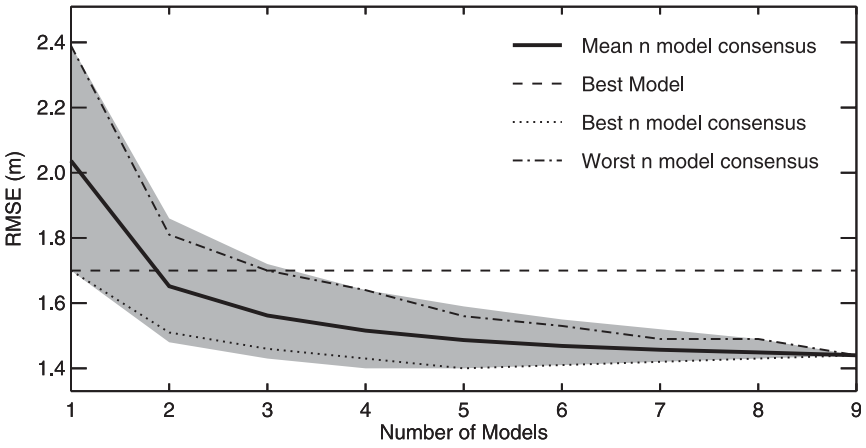

FIG. 3. As in Fig. 2, but for U_{10}.

beyond this could even degrade the forecast. Consider an extreme example, the case where a model predicts a constant 1-m wave height over all time and space. According to the formulation used here, it will receive a nonzero weighting in an objectively constructed composite, despite the fact that it is likely to degrade the forecast. In practice, this degradation only occurs when particularly poor models are added to a set of models where most of the gains to be had through this type of compositing have already been realized.

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

To illustrate this point, consider the simple case of constructing a consensus forecast from model C and one other model. The models chosen here are those with the highest and lowest error correlations with model C, namely models I and A with error correlations of 0.85 and 0.43, respectively (the error correlation coefficients between all raw models are given in the appendix). Statistics for each of these raw models, as well as the resultant PWBC consensus forecasts, are given in Table 9. From the raw model statistics, it is apparent that model I is a better performing model than model A in this case. However, due to the low error correlation between models C and A, the consensus forecast using these models does better than that using the higher quality model I. This explains why the best n models and worst n models do not lie at the extremes of the spread of all possible combinations in Figs. 2 and 3.

c. *Extended forecast periods*

Up until this point, only 24-h forecasts have been considered. The following examines how these corrections perform for extended forecast periods out to 5 days. Of the 10 models used in the previous section, only 6 produce forecasts out to 5 days, namely models A, B, C, E, I, and G. 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 six models for all forecast periods. Bias correction and component weightings are calculated independently for each forecast period, resulting in different weightings for different forecast periods. Hence, if a model's relative performance deteriorates through the forecast period, so too will its weighting in the corresponding composite forecast.

TABLE 9. The 24-h H_s forecast statistics for the raw models C, A, and I, as well as the PWBC results for different combinations of these models. Here, R refers to the error correlation between the models used in the PWBC consensus.

Model	MAE (m)	RMSE (m)	SI
C	0.27	0.38	0.15
A	0.40	0.54	0.20
I	0.33	0.45	0.17

PWBC using specified components

Component models	*R*	MAE (m)	RMSE (m)	SI
C and A	0.43	0.25	0.34	0.13
C and I	0.85	0.27	0.37	0.15

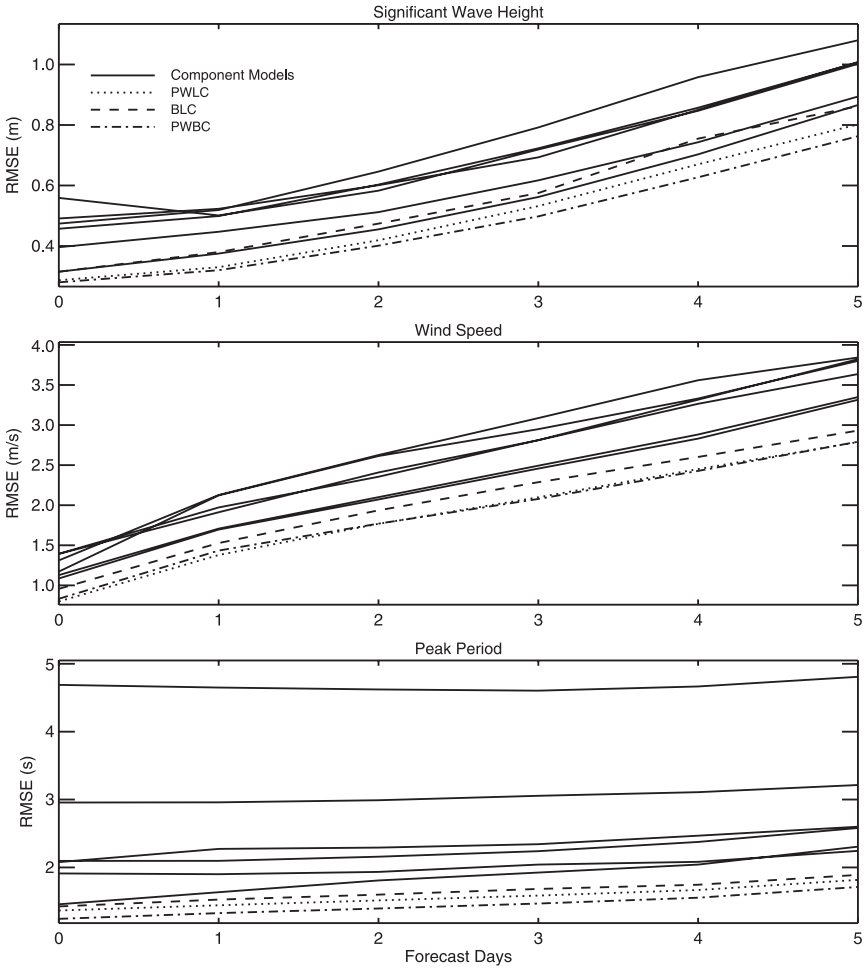

FIG. 4. RMSE growth with forecast period for models for which 5-day forecasts are available (models A, B, C, E, G, and I) as well as PWBC, PWLC, and BLC forecasts produced using these models: (top) significant wave height, (middle) wind speed, and (bottom) peak period.

Figure 4 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, U_{10}, and T_p. Across all models, U_{10} shows a rapid increase in forecast error from analysis to 24-h forecast, steadily increasing throughout the remaining forecast period. For H_s, the error increases only slightly between the analysis and the 24-h forecast. Several factors may explain this difference. For the atmospheric models, the active error reduction during the analysis results in a relative jump in error as the model stabilizes during the forecast period. Four of the six wave models shown here do not assimilate wave data; therefore, this does not apply. In the case of the remaining two that do, 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 or U_{10}. This is likely due, at least in part, to the difficulties in accurately modeling T_p, as discussed in section 5a.

Similar analysis to that presented in section 5a for 24-h forecasts was carried out for all forecast periods. Learned correction schemes appear to retain their advantage throughout the forecast period, with results varying little from 24-h results.

It is also evident from Fig. 4 that, in the case of H_s, the PWBC composite is consistently about 1 day better than the best model; that is, the RMSE of the composite at forecast day n is roughly that of the best individual model at day $n - 1$. For U_{10}, the relative gains are even greater.

6. Summary

The OCF scheme has been applied to forecasts of H_s, U_{10}, and T_p. Forecasts have been compiled using 10 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.

Overall, it is found that the skill levels of the site forecasts of these parameters are considerably improved through the use of OCF techniques. Performance-weighted composite schemes were found to be the best performers, with H_s and T_p achieving the best results using bias-corrected components, and U_{10} using linearly corrected components. These composites resulted in improvements of 36%, 47%, and 31% in RMSE values over the mean raw model components, respectively. Despite the dominance of a single model, composite techniques add value to the forecast. Not only is the forecast skill of the best model improved upon by the addition of further, less skillful models using OCF techniques, but it is possible to beat this model with composites that do not include it. These improvements are found to persist throughout the forecast period out to 5 days, with a 24-h improvement in forecast skill gained.

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 minimization during compositing, the degree of error correlation between models chosen for the composite must be considered in addition to the quality of the model with the aim being to maximize the degree to which component errors are out of phase.

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 these models. It was found that the best correction scheme was a so-called composite of composites, whereby these three models were first averaged and then OCF was 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 exists for OCF forecasts to be extended to grid-based forecasts using altimeter observations.

Acknowledgments. The authors would like to acknowledge the work done by Jean Bidlot at ECMWF in his continued efforts with the intercomparison project that provides the data for this work. We would also like to thank all the institutions that contribute their data to this project, specifically UKMO, FNMOC, MSC, NCEP, METFR, DWD, SHOM, JMA, KMA, and PRTOS. We thank Eric Schulz and Tim Hume, as well as three anonymous reviewers, for their helpful comments on the manuscript.

APPENDIX

Raw Model Error Correlations

TABLE A1. Error correlations between each model for 24-h H_s forecasts. The highest and lowest correlations with model C are shown in boldface (models I and A, respectively).

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H	Model I	Model J
Model A	1.00	0.41	**0.43**	0.40	0.38	0.51	0.51	0.40	0.49	0.23
Model B		1.00	0.54	0.42	0.78	0.22	0.46	0.47	0.59	0.21
Model C			1.00	0.66	0.56	0.57	0.62	0.57	**0.85**	0.45
Model D				1.00	0.43	0.54	0.56	0.49	0.61	0.37
Model E					1.00	0.24	0.41	0.42	0.58	0.23
Model F						1.00	0.58	0.57	0.48	0.45
Model G							1.00	0.75	0.66	0.43
Model H								1.00	0.57	0.46
Model I									1.00	0.30
Model J										1.00

REFERENCES

Bidlot, J. R., and M. Holt, 2006: Verification of operational global and regional wave forecasting systems against measurements from moored buoys. JCOMM Tech. Rep. 30, WMO/TD-1333, JCOMM, 12 pp.

——, 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. *Wea. Forecasting,* **17,** 287–310.

——, and Coauthors, 2007: Inter-comparison of operational wave forecasting systems. *Proc. 10th Int. Workshop on Wave Hindcasting and Forecasting*, Oahu, HI, WMO/IOC Joint Technical Commission for Oceanography and Marine Meteorology, H1.

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

Caires, S., and A. Sterl, 2003: Validation of ocean wind and wave data using triple collocation. *J. Geophys. Res.,* **108,** 3098, doi:10.1029/2002JC001491.

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

Chen, H. S., 2006: Ensemble prediction of ocean waves at NCEP. *Proc. 28th Ocean Engineering Conf. in Taiwan*, Kaohsiung, Taiwan, National Sun Yat-Sen University, 25–37.

Cheng, W. Y. Y., and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting,* **22,** 1304–1318.

Clemen, R., and R. Winkler, 1985: Limits for the precision and value of information from dependent sources. *Oper. Res.,* **33,** 427–442.

Durrant, T., D. Greenslade, and I. Simmonds, 2009: Validation of *Jason-1* and *Envisat* remotely sensed wave heights. *J. Atmos. Oceanic Technol.,* **26,** 123–134.

Efron, B., and R. Tibshirani, 1991: Statistical data analysis in the computer age. *Science,* **253,** 390–395.

Engel, C., and E. Ebert, 2007: Performance of hourly operational consensus forecasts (OCFs) in the Australian region. *Wea. Forecasting,* **22,** 1345–1359.

Farina, L., 2002: On ensemble prediction of ocean waves. *Tellus,* **54A,** 148–158, doi:10.1034/j.1600-0870.2002.01301.x.

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. *J. Appl. Meteor.,* **11,** 1203–1211.

Hibon, M., and T. Evgeniou, 2005: To combine or not to combine: Selecting among forecasts and their combinations. *Int. J. Forecasting,* **21,** 15–24, doi:10.1016/j.ijforecast.2004.05.002.

Hoffschildt, M., J. Bidlot, B. Hansen, and P. Janssen, 2000: Potential benefit of ensemble forecasts for ship routing. ECMWF Tech. Rep. 287, Reading, United Kingdom, 25 pp.

Janssen, P. A. E. M., 2000: Potential benefits of ensemble prediction of waves. ECMWF Newsletter, No. 86, ECMWF, Reading, United Kingdom, 3–6.

——, B. Hansen, and J. R. Bidlot, 1997: Verification of the ECMWF wave forecasting system against buoy and altimeter data. *Wea. Forecasting,* **12,** 763–784.

Komen, G., 1994: *Dynamics and Modelling of Ocean Waves*. Cambridge University Press, 532 pp.

Queffeulou, P., 2004: Long-term validation of wave height measurements from altimeters. *Mar. Geod.,* **27,** 495–510.

Tolman, H. L., 1991: A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents. *J. Phys. Oceanogr.,* **21,** 782–797.

——, 2002: Validation of WAVEWATCH III version 1.15 for a global domain. NCEP Tech. Note 213, 33 pp.

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*, Las Vegas, NV, Amer. Meteor. Soc., 57–62.

Wonnacott, T., and R. Wonnacott, 1972: *Introductory Statistics for Business and Economics*. John Wiley and Sons, 622 pp.

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

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