Consensus of Numerical Model Forecasts of Significant Wave Heights

FRANK WOODCOCK AND DIANA J. M. GREENSLADE

Bureau of Meteorology Research Centre, Melbourne, Victoria, Australia

(Manuscript received 30 May 2006, in final form 18 October 2006)

ABSTRACT

The operational consensus forecast (OCF) scheme uses past performance to bias correct and combine numerical forecasts to produce an improved forecast at locations where recent observations are available. Here, OCF uses past observations and forecasts of significant wave height from five numerical wave models available in real time at the Australian Bureau of Meteorology. In addition to OCF, different adaptive weighting and forecast combination strategies are investigated. At deep-water sites (ocean depth > 25 m), all of the interpolated raw model forecasts outperformed 24-h persistence and, after bias correction, one model was clearly best. Significant improvements over raw model significant wave height forecasts were achieved by bias correction, linear-regression methods, and combination strategies. The best forecasts were obtained from a "composite of composites" in which models with highly correlated errors were combined before being included in the performance-weighted bias-corrected forecast. This technique slightly outperformed the linear-regression-corrected best model. At shallow-water sites (ocean depth < 25 m), all raw models perform poorly relative to the 24-h persistence. The composited, corrected forecasts significantly improved on raw model significant wave height forecasts but only slightly outperformed the 24-h persistence. The raw models generated unrealistically large biases that tended to be amplified with larger observed values of significant wave height.

1. Introduction

The operational consensus forecast (OCF) scheme (Woodcock and Engel 2005, referred to hereafter as WE05) 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 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.

This study employs OCF to generate 24-h predictions of significant wave height at 18 wave observation locations around Australia. Direct model output forecasts, interpolated from numerical wave models (five models for deep-water sites and four for shallow-water sites), provide the underlying component forecasts in the OCF composite. The main objective is to investigate whether OCF improves on its component forecasts. Additionally, several modifications to the WE05 OCF procedure were undertaken to investigate the impacts from variations in training-period and combination strategies.

Typical operational global wave models provide wave forecasts that are skillful only in water depths greater than about 25 m (Booij et al. 1999). This is mainly due to their lack of detailed shallow-water physics and bathymetry (e.g., see Gorman et al. 2003). The observations for this study were obtained from buoys located in both shallow and deep water. The inclusion of shallow-water sites (several being within the complex

Corresponding author address: Frank Woodcock, Bureau of Meteorology Research Centre, P.O. Box 1289 K, Melbourne, VIC 3001, Australia.
E-mail: f.woodcock@bom.gov.au

DOI: 10.1175/WAF1021.1

bathymetry of the Great Barrier Reef) provides an opportunity to investigate the provision of accurate forecasts in shallow water. Additionally, it can provide a benchmark for future evaluations of shallow-water wave models.

The observational data and numerical models are described in section 2, details of the method in section 3, results in section 4, while section 5 contains a summary of the work and some potential avenues for improvement.

2. Data

Observational data and model forecasts used in this study were from 1 November 2003 for buoys in a depth of less than 25 m (i.e., shallow water) and from 16 July 2004 for deep-water buoys until 31 May 2005. These periods differed because one model used for deep water was only available at the bureau after July 2004 and it does not provide shallow-water predictions.

The study was based on operationally available data. As such, there were frequent occasions when model predictions and/or observational data were unavailable. Missing data were far more prevalent than in WE05, which led to a modification from the WE05 scheme to minimize the impact of missing data when comparing OCF with other bias-correction and compositing schemes. Here, the training set consisted of the immediate prior n events, where an event is a single occasion when *all* forecasts and their verifying observations are present. This restriction differs substantially from the operational OCF where depleted component sets and up to half the observations in the learning window are tolerated before OCF is discontinued until sufficient observations are available for learning to resume. The restriction was imposed on the assumption that the more complex regression methods of bias correction and compositing could perform poorly compared to OCF when data were missing from the training period. The results comparing performance with varying training periods (section 4e) justifies this assumption. The intention was to enable an identical treatment of all experimental methods and their verification over matching events.

Bias-correction and compositing parameters were allowed to persist across missing data sequences rather than be relearned whenever the daily sequence of events dropped below the nominated training period number. This change from the operational OCF was implemented both to offset the impact of missing data in the comparison of methods and to permit the use of additional sites with relatively few observations by avoiding any additional verification data loss imposed by relearning.

a. Observations

Observations from the Australian national wave data network were used to bias correct, weight, and verify the forecasts. The locations of the observation sites are shown in Fig. 1 with the details of each instrument listed in Table 1. Note that the locations included 13 deep-water sites and 5 shallow-water sites.

The observations within the Australian national wave data network are predominantly from Waverider buoys. The basis of the Waverider buoy system is a spherical buoy tethered by a mooring to follow the vertical motions of the water surface. Within the buoy an accelerometer is mounted to detect only the vertical movement of the buoy as it rides on the water surface. Directional Waverider buoys detect the horizontal motion of the buoy as well as the vertical movement and hence are able to calculate the direction of the wave motion. The directional information from these buoys however is not used in this study. The laser wave gauge uses a laser mounted above the water surface to measure the surface displacement.

Vertical accelerations from the Waverider buoys are integrated to obtain the surface displacement for a subset of each hour—typically 20–30 min. The time series of surface displacements are then analyzed to produce hourly significant wave heights H_s, where H_s is calculated by using

$$H_s = 4\sqrt{m_0}$$

and m_0 is the variance of the surface displacement time series. Here, H_s is approximately equal to the average of the highest one-third of the waves. The estimated sampling error of the buoy measurements of H_s is 7%–8% of the observed value (Donelan and Pierson 1983; Monaldo 1988).

The last column in Table 1 lists the number of verifiable, independent forecasts using a running 29-event training set. The large variations are due to both different study periods for deep and shallow sites and variations in the availability of the observations.

b. Models

Numerical forecasts from five wave models were considered. These are all available in real time at the bureau. Table 2 lists some of their details.

The first three models are implementations of AUSWAM, a version of the third-generation Wave Model (WAM; WAMDI Group 1988; Komen et al. 1994). All three versions of AUSWAM cover the Australian region with different domain sizes, different spatial resolutions, and different sources of wind forcing. WAMMES is nested within WAMAUS, which in turn

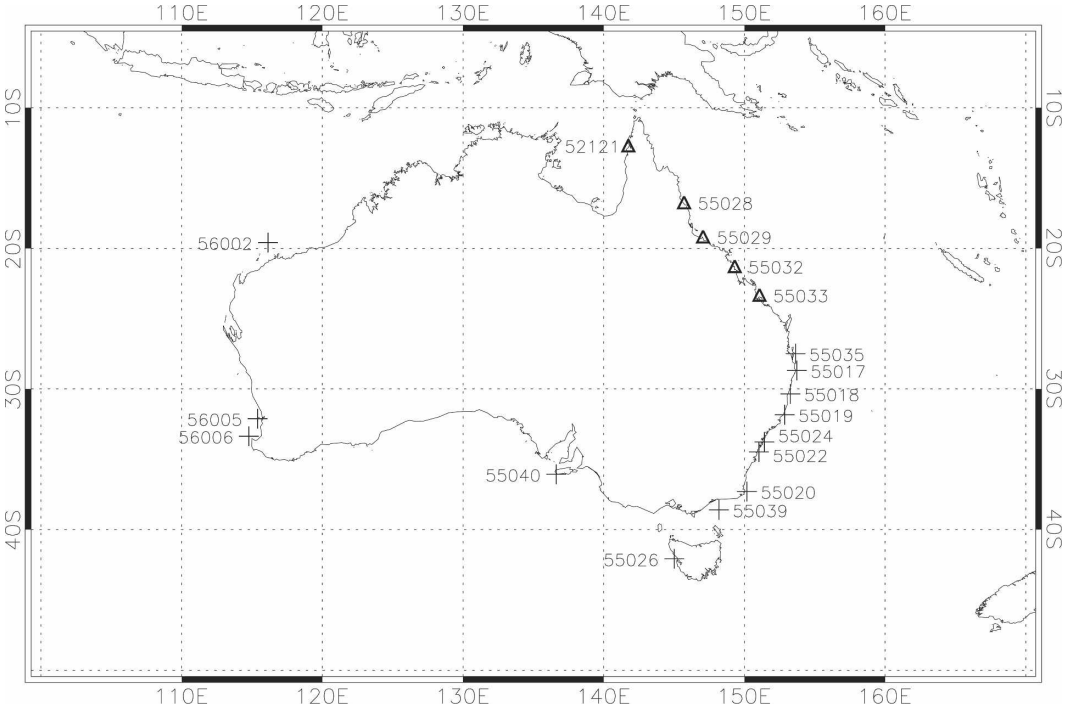

FIG. 1. Location of the observation sites (see Table 1). The shallow-water sites are indicated with triangles.

is nested within WAMGLOB. Specific details of the operational implementations of AUSWAM can be found in National Meteorological and Oceanographic Centre (1999) and Greenslade (2001).

The fourth model used in this work is the wave model run operationally at the Met Office (UKMO). It is a second-generation wave model that includes the assimilation of altimeter wave heights (Holt 1997). The final model used is the operational wave model from the National Centers for Environmental Prediction

TABLE 1. Details of the observation sites used (see Fig. 1). Here, WMO refers to the number assigned to the location by the World Meteorological Organization. Instrument types are Waverider (W), directional Waverider (DW), and Laser Wave Gauge (LWG). Owners of the datasets are the Manly Hydraulics Laboratory (MHL); the Environmental Protection Agency (EPA), Queensland, Australia; Western Australia Department for Planning and Infrastructure (DPI); Woodside Petroleum Ltd. (Wood); Bureau of Meteorology (BoM); and Esso Australia Ltd. (ESSO). Entries in italics refer to the buoys that are designated "shallow water."

WMO	Name	Lat (°S)	Lon (°E)	Depth (m)	Type	Owner	No. of forecasts
52121	*Weipa*	*12.68*	*141.75*	*7*	*W*	*EPA*	*1405*
55017	Byron Bay	28.69	153.73	72	DW	MHL	91
55018	Coffs Harbor	30.35	153.27	73	W	MHL	91
55019	Crowdy Head	31.83	152.86	79	W	MHL	64
55020	Eden	37.29	150.18	110	W	MHL	91
55022	Port Kembla	34.48	151.03	78	W	MHL	86
55024	Sydney	33.77	151.42	85	W	MHL	545
55026	Strahan	42.08	145.01	100	W	BOM	528
55028	*Cairns*	*16.73*	*145.71*	*14*	*W*	*EPA*	*1683*
55029	*Townsville*	*19.16*	*147.06*	*20*	*DW*	*EPA*	*58*
55032	*Hay Point*	*21.27*	*149.31*	*10*	*W*	*EPA*	*92*
55033	*Emu Park*	*23.31*	*151.07*	*22*	*DW*	*EPA*	*1526*
55035	Brisbane	27.49	153.63	70	DW	EPA	568
55039	Kingfish B	38.60	148.19	78	LWG	ESSO	578
55040	Cape du Couedic	36.07	136.62	80	W	BoM	538
56002	North Rankin	19.59	116.14	125	W	Wood	447
56005	Rottnest	32.11	115.40	48	W	DPI	90
56006	Cape Naturaliste	33.36	114.78	50	W	DPI	522

TABLE 2. Characteristics of wave models.

Model	Center	Domain	Spatial resolution	Wind forcing	Data assimilation?	Shallow water?
WAMGLOB	BoM	Global	1°	3-hourly	Yes	No
WAMAUS	BoM	Regional	1/2°	1-hourly	Yes	No
WAMMES	BoM	Regional	1/8°	1-hourly	No	Yes
UKMO	Met Office	Global	5/6° × 5/9°	1-hourly	Yes	Yes
WWIII	NOAA	Global	5/4° × 1°	3-hourly	No	Yes

(NCEP). This model is the WAVEWATCH III (WWIII; Tolman 1991): a third-generation wave model developed at NCEP and similar to the WAM model. WWIII, however, differs in areas such as the model structure, the numerical methods, and the physical parameterizations. The global version of WWIII operates only to a minimum depth of 25 m, so shallow-water predictions of H_s derived from WWIII were unavailable.

The models that include shallow-water physics (see Table 2) generally do so in only a limited way. In particular, the only shallow-water effects explicitly included in WAMMES are the dissipation of energy due to bottom friction and the alteration of the dispersion relation to depend on depth, which results in modifications to the source terms. WWIII and the UKMO model include these effects and also refraction and straining of the wave field.

It is important to note that, while the basic physics of some of the models are similar (e.g., WAMGLOB and WAMAUS), all five models have different wind forcings and spatial resolutions, while some include data assimilation and some include shallow-water physics. These different configurations generate errors that vary between the models and thereby enhance the likelihood of improved forecasts from a consensus of bias-corrected model forecasts—the success of compositing techniques depends in part upon the extent to which these errors are random and out of phase.

3. Method

The 24-h model forecasts of H_s were generated at the observation sites every 12 h by cubic spline interpolation from the nearest model grid points to the observation location. The verifying observation was the closest (in time) observation within an hour of the forecast.

It should be noted that observations of H_s are averages in time at individual locations while the model forecasts are expected values of H_s over an area (the model grid box) and a time period (the model time step). So all estimates of H_s used here, including the different model forecasts, represent different spatial and temporal scales.

Following the OCF methodology, bias correction is undertaken on a training set of events for each component model contributing to the final forecast. The training set bias is estimated from the component errors using the best easy systematic estimator (BES; Wonnacott and Wonnacott 1972, their section 7.3), where

$$\text{BES} = (Q_1 + 2Q_2 + Q_3)/4,$$

where Q_1 and Q_3 are the first and third quartiles of the training set and Q_2 is its median. To simplify our computations here, training sets were restricted to $E + 1$ members where E is a multiple of 4. Hence, in the experiments, training windows were allowed to vary from 5 to 57 events.

The 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 (combining forecasts) 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. Procedures such as OCF combine forecasts using weights according to the MAE of the bias-corrected component forecasts. Here, we follow Daley (1991) and use error variance weights in performance-weighted compositing. In practice, the differences resulting from the use of error variance weights instead of MAE weights is negligible. A short time series of the performance-weighted bias-corrected composite forecast (i.e., OCF) and the raw model forecasts for site 55026 (Strahan) is provided in Fig. 2: the WAMGLOB and WAMAUS forecasts were withheld from the plot for clarity.

One external method wherein all component forecasts are treated simultaneously was used. A multilinear-regression equation was generated from the train-

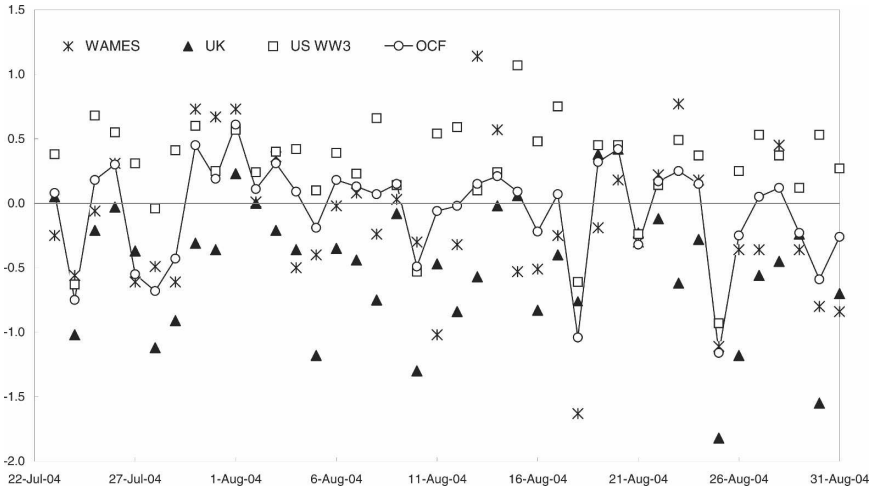

FIG. 2. Time series of raw model and OCF composite errors (m) for site 55026. Raw WAMGLOB and WAMAUS errors were withheld from the plot for clarity. The line linking OCF errors has also been included for clarity.

ing set for all models: a procedure somewhat akin to a running multimodel MOS with the derived H_s as model predictors and the observed H_s as the predictand. In an operational weather prediction scheme servicing several weather elements per site over hundreds of sites, a running multimodel, multilinear-regression option may not be viable. However, here we only have one element (H_s), 13 sites, and only four or five predictors.

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 and intercept values change for every forecast).

In summary, the following comparisons were undertaken:

- training-window variations in steps of 4 for 5 to 57 verifiable events;
- internal linear-regression correction and bias-correction forecasts;
- equal-weighted and performance-weighted composites;
- multimodel, multilinear-regression compositing (MM); and
- best linear-regression-corrected component forecast (BELC).

All bias-correction, linear-regression-correction, MM, and BELC comparisons are undertaken over corresponding, matching events (i.e., identical verifying sets and training windows). Matching 24-h persistence forecasts were generated as a benchmark.

4. Results

In this section, only the results from runs using a running 29-event training period are presented in detail. They represent the sequence-independent forecasts (i.e., forecasts following the training sets). The impact of varying training window size is discussed in section 4e. All 29-event results cover exactly the same events. Summary statistics were generated for each site and then combined in Tables 3 and 4.

Verification statistics include bias, MAE, RMSE [median and its 90% confidence interval generated using the bootstrap method (Efron and Tibshirani 1991) with 10^6 iterations], XAE, scatter index (SI% = 100 × standard deviation normalized by the observation mean), and percentage of explained variance (V% = 100 × square of the correlation between forecast and observation). Statistics were calculated separately at each site and then consolidated using event frequency weightings.

a. Deep water

Table 3a summarizes the performance of the interpolated raw model forecasts. These results are calculated based on 4239 independent forecasts over 13 deep-water sites for the 29-event training period. WWIII significantly (more than 95% level based on RMSE) outperformed all of the other models. It yielded the lowest MAE, RMSE, XAE, and SI%, as well as the largest V%. Nevertheless, there was considerable variation across sites with every raw forecast scheme outperforming the others for at least one site.

TABLE 3a. Verification of interpolated raw model 24-h forecasts of significant wave height at 13 deep-water sites over 4239 forecasts per model. Here, SI is the percentage (standard deviation of forecast errors)/(mean of observations). Bias, MAE, RMSE, and maximum absolute error (XAE) are in m. In addition, V% is the percentage of variance in the observed significant wave height explained by the forecasts. The 90% confidence intervals for the overall best raw model forecasts RMSE are included. Persistence is the 24-h persistence forecast. The best results for the table are in boldface.

			RMSE					
Model	Bias (m)	MAE (m)	Median (m)	5%	95%	XAE	SI%	V%
UKMO	−0.02	0.47	0.64	0.62	0.66	2.83	33	52
WWIII	0.14	**0.36**	**0.53**	**0.50**	**0.55**	**2.77**	**27**	**70**
WAMGLOB	0.07	0.43	0.59	0.57	0.62	2.92	32	51
WAMAUS	0.08	0.43	0.59	0.57	0.62	2.80	33	57
WAMMES	−0.01	0.39	0.56	0.54	0.59	2.87	30	62
Persistence	0.00	0.55	0.80	0.77	0.83	3.89	41	52

Component model biases exceeding 0.5 m occurred at some sites. However, all raw model forecasts produced more favorable error statistics than did persistence.

Table 3b summarizes the performance over the same events as Table 3a but with the forecast bias over the previous 29 events removed. Bias correction improved on the raw forecast results by approximately 10%–15%. The bias-corrected WWIII clearly outperforms the other bias-corrected models. This result contrasts markedly with WE05 where all of the bias-corrected model forecasts showed similar accuracy.

Equal-weighted and performance-weighted composites of the bias-corrected forecasts are included in Table 3b. The OCF strategy (i.e., performance-weighted bias correction) yields approximately a 15% improvement on the raw WWIII forecasts. Bias correction alone improved WWIII by 13% so that compositing was not the dominant factor in forecast improvement. OCF marginally outperformed equal-weighted bias correction, so it could be noted that even though WWIII was the best individual corrected model, the inclusion of other models in the compositing provided some further small improvement.

The impact of linear-regression correction on the raw model forecasts is shown in Table 3c. For the U.K. and Australian models the improvement due to linear-regression correction and bias correction was similar. However, WWIII improved more from linear-regression correction than from bias correction. It significantly outperformed the other corrected models. The impact of linear regression increased the gap in performance between WWIII and the other models, and hence compositing was less beneficial than with bias correction. In fact, neither equal-weighted nor performance-weighted composites outperform the linear-regression-corrected WWIII. The multimodel linear regression (MM in Table 3c) marginally outperformed the internal linear-regression schemes and was especially useful in reducing the XAE.

Applying linear regression to the best of the raw forecasts in the training period and using the regression equation on the next corresponding forecast (BELC) produced forecasts that matched the linear-regression-corrected WWIII and the composites for accuracy.

Table 3d compares the performance of the experimental results with the average of the five raw models. Composite forecast statistics improved the average raw model forecast statistics by between 15% and 30%. At 9 out of the 13 deep-water sites performance-weighting methods including MM (3 sites) produced the lowest

TABLE 3b. As in Table 3a but applying a bias correction developed from the training window. Here, EW BC refers to the equal-weighted bias-corrected composite and OCF refers to the performance-weighted bias-corrected composite.

			RMSE					
Model	Bias (m)	MAE (m)	Median (m)	5%	95%	XAE	SI%	V%
UKMO	−0.02	0.40	0.55	0.53	0.57	2.79	29	55
WWIII	0.00	0.31	0.46	0.43	0.49	2.69	24	**72**
WAMGLOB	−0.02	0.38	0.53	0.51	0.56	3.01	28	55
WAMAUS	−0.01	0.36	0.52	0.49	0.54	2.81	28	59
WAMMES	0.00	0.34	0.51	0.49	0.54	2.72	26	61
EW BC	−0.01	0.30	0.44	0.41	0.47	**2.48**	**23**	70
OCF	0.00	**0.29**	**0.42**	**0.40**	**0.45**	2.59	**23**	70

TABLE 3c. As in Table 3b but applying linear-regression corrections derived from the training window. Here, EW LC refers to the equal-weighted linear-regression-corrected composite, PW LC is the performance-weighted linear-regression-corrected composite, BELC is the best linear-regression-corrected component forecast (see text for details), and MM is a multimodel linear-regression composite.

Model	Bias (m)	MAE (m)	RMSE			XAE	SI%	V%
			Median (m)	5%	95%			
UKMO	−0.02	0.39	0.55	0.53	0.57	2.81	29	53
WWIII	0.01	**0.29**	**0.42**	0.40	0.45	2.60	**22**	**72**
WAMGLOB	0.00	0.38	0.54	0.51	0.56	3.15	28	54
WAMAUS	0.00	0.36	0.51	0.48	0.53	2.93	27	58
WAMMES	0.00	0.34	0.49	0.46	0.51	2.96	26	61
BELC*	0.00	0.31	0.43			2.54	23	69
EW LC	0.00	0.31	0.44	0.42	0.47	2.66	23	68
PW LC	0.02	0.32	0.43	0.41	0.46	2.61	24	70
MM	0.01	0.30	0.43	0.41	0.46	**2.54**	**22**	69

* The mean not the median value.

RMSE. At one site (55039), the raw model forecasts from WWIII recorded the lowest RMSE of all. Generally, performance-weighted linear-regression correction outperformed both the best raw forecast and the best bias-corrected forecast scheme. However, the best linear-regression-corrected model was slightly better than the performance-weighted linear-regression-corrected composite and as good as OCF.

If most weight is placed on the RMSE, then the results for deep-water sites indicate the following.

- One raw model (WWIII) significantly outperformed the others and did so after bias correction and linear-regression corrections were applied to all.
- Learned correction strategies (either bias correction or linear-regression correction) substantially improved upon the raw forecasts.
- Linear-regression-correction methods performed similarly to bias-correction methods.
- Compositing of corrected forecasts substantially improved on the average bias-corrected error.
- If anything, performance-weighting composites methods slightly outperformed equal-weighting composites.
- In hindsight, the linearly corrected best model was as good as the best composite. This is discussed further in section 4d.

b. Shallow water

WWIII model forecasts were unavailable for the shallow-water sites so only four numerical models were used. This permitted more events (19 months for shallow water as opposed to 10.5 months for deep water) to be examined.

TABLE 3d. Improvement impacts of bias correction (BC), linear-regression correction (LC), and equal-weighted (EW) and performance-weighted (PW) compositing over the average raw model for deep-water events. BELC and MM are as in Table 3c. The 24-h persistence is included. "Best" refers to the best overall single model, i.e., hindsight selection, and for raw, BC, and LC it was WWIII.

Model	MAE (m)	RMSE (m)	XAE	SI%	V%
Avg raw	0.42	0.58	2.84	31.0	58.4
	Improvement over avg of raw models (%)				
Best raw	13	9	2	13	20
Avg BC	14	12	1	13	3
Avg LC	15	14	−2	15	2
Best BC	25	21	5	23	**23**
Best LC	**30**	**28**	8	**29**	**23**
PW BC (i.e., OCF)	**30**	**28**	9	26	20
EW BC	28	24	**13**	26	20
PW LC	23	26	8	23	20
EW LC	25	24	6	26	16
MM	28	26	11	**29**	18
BELC	25	26	11	26	18
Persistence	−32	−37	−37	−32	−11

TABLE 4a. As in Table 3a but for five shallow-water sites and 4851 forecasts.

Model	Bias (m)	MAE (m)	RMSE			XAE	SI%	V%
			Median (m)	5%	95%			
UKMO	0.54	0.57	0.72	0.70	0.74	3.42	79	49
WAMGLOB	0.53	0.56	0.67	0.65	0.70	3.37	73	41
WAMAUS	0.48	0.52	0.64	0.62	0.65	3.62	72	48
WAMMES	0.23	0.43	0.56	0.53	0.59	3.75	72	**53**
Persistence	0.00	**0.17**	**0.28**	0.26	0.33	**3.06**	**48**	51

Comparing the corresponding raw forecast results in Table 3a (deep water) with Table 4a (shallow water) reveals that model biases in shallow water are a significant problem. The three coarser-resolution models have biases of the order of 0.5 m while the WAMMES bias is half that. WAMMES, which includes bottom friction, a depth-dependent dispersion relation, and higher spatial resolution, significantly outperformed the other models. Nevertheless, the 24-h persistence forecasts overall were significantly better than any model forecasts.

The poor performance of the models in shallow water compared to deep water is probably due to the absence of both shallow-water physics and to their relatively coarse resolution. The shallow-water buoys are generally closer to the coastline and effects due to complex reef topography and bathymetry, such as sheltering, are more important. Coarse-resolution models will not be able to capture these dynamics accurately. Figure 3 illustrates a large bias in the raw model forecasts that often exceeds the observed H_s and which appears to be amplified whenever the observed H_s increases.

The reason for including shallow-water sites in this work is to examine whether it is possible to make simple corrections to accommodate shallow-water effects. For example, a simple linear regression may be appropriate for adjusting the offshore wave height as it approaches the shore (see, e.g., Hemer and Bye 1999).

Not surprisingly, the impact of bias correction was substantial, as Table 4b shows. All of the bias-corrected model forecasts significantly outperformed their raw counterpart and even the best raw model. The bias-correction error statistics indicated about a 45% improvement over the raw models. After bias correction all model performances are similar, which is a corresponding result to WE05. Performance-weighted and equal-weighted composites improved on the individual bias-corrected models by 15% and performed slightly worse than did the 24-h persistence. OCF again slightly outperformed the equal-weighted bias correction. However, note that the SI% and V% statistics are poor compared to their deep-water equivalents. For operational forecasts, a value of SI = 30% shows reasonable skill. This means that even though the RMSE is low compared to deep water, this is an artifact of the lower H_s in shallow water and the forecast skill here is still poor.

Table 4c shows that the linear-regression correction markedly improves over the bias correction in shallow sites. The linear-regression and MM composites produced very similar results and were about 20% better than the bias-corrected composites (probably reflecting the amplification of bias whenever H_s increases, as seen in Fig. 3) and outperformed the 24-h persistence. However, as V% indicates, only a little more than half of the variance in H_s was explained by the corrected model forecasts, and the high SI% values again indicate that the forecast skill is poor.

Table 4d summarizes the shallow-water site results showing the relative performance of the various corrections and composites against the average of the raw

TABLE 4b. As in Table 3b but for five shallow-water sites.

Model	Bias (m)	MAE (m)	RMSE			XAE	SI%	V%
			Median (m)	5%	95%			
UKMO	0.01	0.23	0.36	0.33	0.40	3.05	72	48
WAMGLOB	0.00	0.23	0.35	0.33	0.41	3.04	71	40
WAMAUS	0.01	0.21	0.35	0.32	0.41	3.21	67	46
WAMMES	0.01	0.20	0.36	0.32	0.41	3.41	**65**	50
EW BC	0.01	0.18	0.31	0.28	0.38	3.08	66	52
OCF	0.01	**0.17**	**0.30**	0.28	0.36	**3.00**	66	**53**

TABLE 4c. As in Table 3a but for shallow-water sites.

Model	Bias (m)	MAE (m)	RMSE			XAE	SI%	*V*%
			Median (m)	5%	95%			
UKMO	0.01	0.16	0.31	0.29	0.34	2.53	56	50
WAMGLOB	0.01	0.17	0.30	0.30	0.35	2.52	59	43
WAMAUS	0.00	0.15	0.29	0.28	0.33	2.68	58	49
WAMMES	0.01	**0.14**	0.27	0.26	0.32	2.43	57	**54**
BELC*	0.02	0.15	0.28			2.53	56	50
EW LC	0.01	**0.14**	**0.26**	0.25	0.33	**2.28**	56	**54**
PW LC	0.01	**0.14**	**0.26**	0.24	0.32	2.33	**53**	53
MM	−0.01	**0.14**	**0.26**	0.25	0.32	2.72	58	**54**

* The mean not the median value.

numerical models. Essentially, all of the linear-regression composites, including MM, improved over persistence.

To summarize, for shallow-water sites the following was found.

- The 24-h persistence forecasts outperformed all of the raw model forecasts.
- One raw model was significantly better than the others, but after bias correction and linear-regression correction, they all performed similarly.
- Learned correction strategies (either bias correction or linear regression) significantly improved upon the corresponding raw forecasts. However, the linear-regression improvements were substantially more than bias-correction improvements.
- Compositing of corrected forecasts substantially improved on the average bias-corrected error.
- Performance-weighting composites of bias-corrected forecasts slightly outperformed equally weighted composites but no difference was evident for linear-regression-corrected forecast composites.
- Linear-regression composites outperformed persistence.
- In hindsight, the corrected best model was as good as the best composite.
- None of the methods produced a forecast that was acceptable for operational forecasting (using a 30% threshold for SI).

These results suggest that in order to improve forecasts of H_s in shallow water, specialized shallow-water wave models are necessary.

c. *Correction methodology*

Both shallow- and deep-water site results provide strong evidence that simple running corrections to model forecasts can result in significant forecast improvements. This result is consistent with several other studies in numerical weather prediction, for example, Mao et al. (1999) using multivariate linear regression to predict numerical forecast error and Stensrud and Skindlov (1996), Stensrud and Yussouf (2003, 2005), and WE05 using bias correction.

TABLE 4d. As in Table 3d but for five shallow-water sites. Best for raw, BC, and LC was WAMMES.

Model	MAE (m)	RMSE (m)	XAE	SI%	*V*%
Avg raw	0.53	0.65	3.54	74.0	47.8
	Improvement over avg of raw models (%)				
Best raw	19	14	−6	3	11
Avg BC	59	45	10	7	−4
Avg LC	71	55	28	22	3
Best BC	62	44	4	12	5
Best LC	**74**	58	31	23	**13**
PW BC (i.e., OCF)	68	54	15	11	11
EW BC	66	52	13	11	9
PW LC	**74**	**60**	34	28	11
EW LC	**74**	**60**	**36**	24	**13**
MM	**74**	**60**	23	22	**13**
BELC	72	57	29	24	5
Persistence	68	57	14	**35**	7

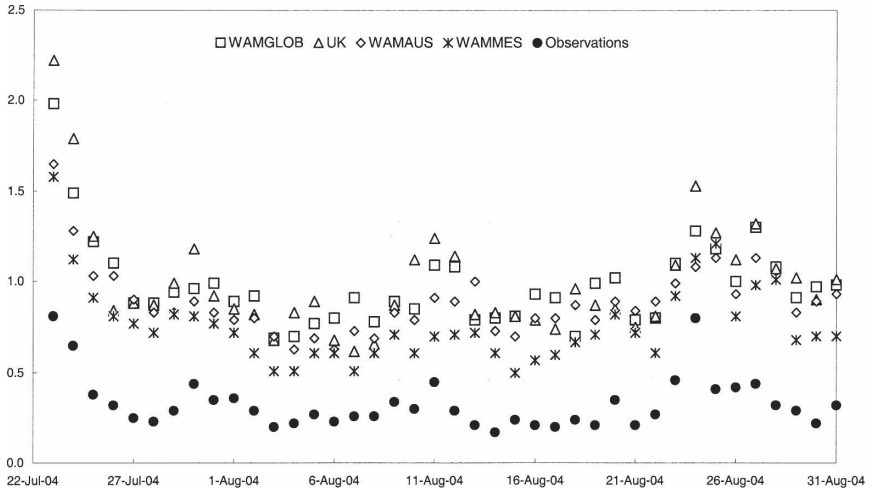

FIG. 3. Raw model forecasts and observed significant wave heights (H_s; m) at the shallow-water site 55028 (Cairns). Note that the model forecasts are often more than twice the observed value and the errors are accentuated whenever the observed H_s increases.

Tables 3b and 3c compare the bias-correction and linear-regression-correction impacts on individual model forecasts of H_s at deep-water sites. Significant (5%) improvements (in RMSE) of linear-regression correction over bias correction were evident for the two better models (WWIII and WAMMES). A slight improvement was evident for WAMAUS and no change was detectable otherwise. For shallow-water sites (Tables 4c and 4d), linear-regression correction significantly (95% level) outperformed bias correction for each model by about 20% on average.

For both shallow- and deep-water sites, linear-regression correction provided significant improvement over bias correction for the best individual models.

d. Compositing methodology

The equal-weighted bias correction and OCF in Tables 3b and 4b agree with the results from WE05 indicating that performance-weighted composites slightly outperform equal-weighted composites. However, there was no evidence that performance weighting outperforms equal weighting when a linear-regression correction is employed (see Tables 3c and 4c).

For deep-water sites OCF provided the most successful composite forecast, whereas at shallow sites, linear-regression-corrected composites performed best. Both internal compositing methods slightly outperformed MM in both deep and shallow water.

The BELC experiment whereby the regression coefficient and intercept from the best linear-regression-corrected component within the training set were applied to the next forecast from that component shows that the method performs almost as well as the best overall linear-regression-corrected model. This is evident by comparing BELC with WWIII in Table 3c for deep water and BELC with WAMMES in Table 4c for shallow water.

A common result in both deep and shallow water was that the linear-regression-corrected best-in-hindsight forecast model performed close to the best of the composite forecasts without the benefit of hindsight. This result may suggest that compositing could be avoided, but as Hibon and Evgeniou (2005) suggest, this hindsight information may not be of much practical value in operations.

Although WAMGLOB, WAMAUS, and WAMMES all have different spatial resolutions and different wind forcing fields, the nesting procedure raises the possibility that their errors would be highly correlated and in phase. Therefore, there is the potential to improve the OCF results by using only one of these models. This possibility was investigated. The average pairwise correlation of errors for the three Australian models was 0.78 while the remaining pairwise correlations averaged 0.58. Thus, there is considerable redundancy of information among the Australian models.

The impact of error cancellation due to the compositing of component model forecasts can be investigated using the difference between the event average of the individual model bias-corrected absolute errors and the absolute value of the corresponding average bias-corrected error. If all of the errors have the same sign, the difference between these averages is zero, whereas if cancellation had maximum impact, then the differ-

TABLE 5. As in Table 3a but a comparison of a composite from (WAMGLOB + WAMAUS + WAMMES)/3, and U.K. and WWIII models against OCF (as in Table 3b) and against the linear-corrected WWIII (as in Table 3c).

Model	Bias (m)	MAE (m)	RMSE (m)	XAE	SI%	*V*%
OCF (composite)	−0.01	**0.28**	**0.40***	**2.47**	**22**	**73**
OCF	0.00	0.29	0.42	2.59	23	70
WWIII	0.01	0.29	0.42	2.60	22	72

* The mean not the median value.

ence would be the average of the absolute errors. Normalizing the outcomes by the average of the absolute errors produces a cancellation index between 0% when there is no benefit from error cancellation and 100% for the maximum benefit of cancellation. The index can then be averaged over all events to give a measure of which combinations are most useful. The average cancellation index for the three pairwise Australian model combinations was 6.4% compared to an average of 12.8% for the remaining combination of pair.

Hence, both cross-correlation and cancellation impact considerations suggest that a better result could be achieved by using only one of the Australian models, or better, by consolidating them. A comparison of the resulting OCF (consolidated Australian models, the UKMO model, and the WWIII), the original OCF (Table 3b), and the linearly corrected WWIII (Table 3c) is provided in Table 5. As expected, the consolidated OCF slightly outperforms the original OCF. It also outperforms the linearly corrected WWIII.

e. Training period

The impact of different training periods on the compositing method performance was consistent across deep- and shallow-water sites. Bias-correction methods stabilized by five or nine events with only negligible changes in bias, MAE, RMSE, SI, or V with further increases in training set size. The errors of the linear-regression-corrected composites decreased more slowly as the training set increased but had stabilized by 13 events. As would be expected, MM performed poorly with a small training set but stabilized by 17 events and continued to exhibit very small improvements as the training set size increased thereafter. The results for bias correction agree with those of Stensrud and Yussouf (2005) and WE05.

5. Summary and further work

The OCF scheme has been applied to forecasts of H_s, in both deep and shallow water, at locations where observations of H_s are available. Our shallow-water results differed markedly from those in deep water. Four of the shallow-water sites are within the complex bathymetry of the Great Barrier Reef, where models have insufficient spatial and bathymetric resolution and shallow-water physics needed for the task. Day-to-day variations in H_s in shallow water are small compared to deep water; consequently, the 24-h persistence proved a far better predictor than the wave models in shallow water. The best model in shallow water was the highest-resolution model, WAMMES, but as the bias correction shows this was due to its smaller bias. Once bias was removed, the models performed similarly in capturing about 50% of the daily variation in H_s. This result suggests that to capture the dynamics of shallow-water waves accurately, a specialized shallow-water wave model, such as the Simulating Waves Nearshore model (SWAN; Booij et al. 1999) may be needed.

In deep water, the models performed far better than persistence. The best forecasts were obtained from a "composite of composites" in which models with highly correlated errors were combined before being included in the performance-weighted bias-corrected forecast. This technique slightly outperformed the linear-regression-corrected best model.

The broad conclusions are that

- in deep water, a 20%–30% improvement over model forecasts of H_s can be achieved using the OCF strategy of performance-weighted compositing of bias-corrected model forecasts; and
- in shallow water, the strategy of compositing model forecasts after linear correction can yield a 60%–70% improvement over raw model forecasts of H_s.

Multimodel, multilinear regression, the most complex of the correction and compositing schemes, was as good as other correction and composite strategies used, but no better than the simpler strategies.

Potential for improvement

There are various avenues that could be pursued to improve these results. One option is to include more and better wave models in the compositing. The present study was limited to those models that were

available in real time at the bureau; however, there are numerous forecasting centers that operate global wave models (see, e.g., Bidlot et al. 2002) and further benefits might be expected with the use of models from other centers.

Further work could improve the use of observational data, for example, by smoothing the hourly buoy data over several hours to represent model time scales better. This study has not addressed forecasts of wave period but could be extended to do so. The inherent noisiness of the peak-period data, however, suggests that it would be difficult to obtain significant improvements in forecasts of peak period through compositing. Mean wave period could be a viable alternative, although it is less commonly used as a forecast product. The extension of the technique to use wave direction data (if available) in a multivariate scheme might improve the results at the shallow-water sites.

There is ongoing research at the bureau to extend the site-based NWP OCF scheme to a grid-based scheme and thus provide corrected forecasts at every grid point in a domain instead of just those locations at which the observations are routinely available. If multiple surface wind fields were available, the OCF technique could be applied to obtain the best possible forecasts of surface marine winds, perhaps using scatterometer data as the verifying observations. The resulting surface wind forecasts could then be used to force a wave model. Under the assumption that most of the error in wave forecasts arises from errors in the surface winds (in deep water at least), this should lead to improved predictions of ocean waves.

There are also connections here with data assimilation methods that could be explored further. For example, the corrections based on the training sets could feed into estimates of the magnitude of the model prediction error that are required for data assimilation schemes. Indeed, the relationship between the corrections at neighboring sites could shed some light on the spatial scales of the model error, another little known but essential component of data assimilation schemes (see, e.g., Greenslade and Young 2004).

Acknowledgments. Thanks to Beth Ebert and Eric Schulz for their helpful and constructive reviews of the draft manuscript and to the three anonymous external reviewers for many helpful suggestions.

REFERENCES

Bidlot, J.-R., D. J. Holmes, P. A. Wittman, 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.

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

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

Donelan, M., and W. J. Pierson, 1983: The sampling variability of estimates of spectra of wind-generated waves. *J. Geophys. Res.,* **88,** 4381–4392.

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

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Greenslade, D. J. M., 2001: The assimilation of *ERS-2* significant wave height data in the Australian region. *J. Mar. Syst.,* **28,** 141–160.

——, and I. R. Young, 2004: Background errors in a global wave model determined form altimeter data. *J. Geophys. Res.,* **109,** C09007, doi:10.1029/2004JC002324.

Gorman, R. M., K. R. Bryan, and A. K. Liang, 2003: Wave hindcast for the New Zealand region: Nearshore validation and coastal wave climate. *N. Z. J. Mar. Freshwater Res.,* **37,** 567–588.

Hemer, M. A., and J. A. T. Bye, 1999: The swell climate of the south Australian sea. *Trans. Royal Soc. S. Aust.,* **123** (3), 107–113.

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

Holt, M. W., 1997: Assimilation of ERS-2 altimeter observations into a global wave model. *Research Activities in Atmospheric and Oceanic Modelling—1997,* WGNE Rep. 25, WMO/TD-792, 8.31–8.32.

Komen, G. J., L. Cavaleri, M. Donelan, K. Hasselmann, S. Hasselmann, and P. A. E. M. Janssen, 1994: *Dynamics and Modelling of Ocean Waves.* Cambridge University Press, 532 pp.

Mao, Q., R. T. McNider, S. F. Mueller, and H. H. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting,* **14,** 190–202.

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

National Meteorological and Oceanographic Centre, 1999: Changes to the operational sea state forecast system. Bureau of Meteorology Operations Bull. 47, Melbourne, Australia, 6 pp.

Stensrud, D. J., and J. A. Skindlov, 1996: Gridpoint predictions of high temperature from a mesoscale model. *Wea. Forecasting,* **11,** 103–110.

——, and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131,** 2510–2524.

——, and ——, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.,* **12,** 217–230.

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.

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

Wonnacott, T. H., and R. J. Wonnacott, 1972: *Introductory Statistics*. Wiley, 510 pp.

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